Stochastic Primal-Dual Hybrid Gradient Algorithm with Arbitrary Sampling and Imaging Applications

SIAM J. OPTIM.
Vol. 28, No. 4, pp. 2783–2808

STOCHASTIC PRIMAL-DUAL HYBRID GRADIENT ALGORITHM WITH ARBITRARY SAMPLING AND IMAGING APPLICATIONS*

ANTONIN CHAMBOLLE†, MATTHIAS J. EHRHARDT‡, PETER RICHTÁRIK§, AND CAROLA-BIBIANE SCHÖNLIEB‡

Abstract. We propose a stochastic extension of the primal-dual hybrid gradient algorithm studied by Chambolle and Pock in 2011 to solve saddle point problems that are separable in the dual variable. The analysis is carried out for general convex-concave saddle point problems and problems that are either partially smooth / strongly convex or fully smooth / strongly convex. We perform the analysis for arbitrary samplings of dual variables, and we obtain known deterministic results as a special case. Several variants of our stochastic method significantly outperform the deterministic variant on a variety of imaging tasks.

Key words. convex optimization, primal-dual algorithms, saddle point problems, stochastic optimization, imaging

AMS subject classifications. 65D18, 65K10, 74S60, 90C25, 90C15, 92C55, 94A08

DOI. 10.1137/17M1134834

1. Introduction.

Many modern problems in a variety of disciplines (imaging, machine learning, statistics, etc.) can be formulated as convex optimization problems. Instead of solving the optimization problems directly, it is often advantageous to reformulate the problem as a saddle point problem. A very popular algorithm to solve such saddle point problems is the primal-dual hybrid gradient (PDHG)[1] algorithm [37, 21, 13, 36, 14, 15]. It has been used to solve a vast amount of state-of-the-art problems—to name a few examples in imaging: image denoising with the structure tensor [22], total generalized variation denoising [11], dynamic regularization [7], multimodal medical imaging [27], multispectral medical imaging [43], computation of nonlinear eigenfunctions [26], and regularization with directional total generalized

*Received by the editors June 16, 2017; accepted for publication (in revised form) August 6, 2018; published electronically October 2, 2018.

http://www.siam.org/journals/siopt/28-4/M113483.html

Funding: The work of the first author was supported by the ANR, "EANOI" project I1148 / ANR-12-IS01-0003 (joint with FWF); part of this work was done while he was hosted in Churchill College and DAMTP, Centre for Mathematical Sciences, University of Cambridge, thanks to support from the French Embassy in the UK and the Cantab Capital Institute for the Mathematics of Information. The work of the second and fourth authors was supported by Leverhulme Trust project "Breaking the non-convexity barrier," EPSRC grant EP/M00483X/1, EPSRC centre grant EP/N014588/1, the Cantab Capital Institute for the Mathematics of Information, and from CHiPS (Horizon 2020 RISE project grant). The second author carried out initial work supported by the EPSRC platform grant EP/M020533/1. Moreover, the fourth author is thankful for support by The Alan Turing Institute. The work of the third author was supported by EPSRC Fellowship in Mathematical Sciences grant EP/N005538/1, entitled "Randomized algorithms for extreme convex optimization."

†CMAP, CNRS, Ecole Polytechnique, Palaiseau 91128, France (antonin.chambolle@cmap.polytechnique.fr).

‡Department for Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, UK (m.j.ehrhardt@damtp.cam.ac.uk, cbs31@cam.ac.uk).

§Visual Computing Center & Extreme Computing Research Center, KAUST, Thuwal 23955, Saudi Arabia, School of Mathematics, University of Edinburgh, Edinburgh EH9 3PD, UK, and The Alan Turing Institute, London NW1 2DB, UK (peter.richtarik@kaust.edu.sa, peter.richtarik@ed.ac.uk).

[1]We follow the terminology of [14] and call the algorithm simply PDHG. It corresponds to PDHGMu and PDHGMp in [21].

variation [29]. Its popularity stems from two facts: First, it is very simple and therefore easy to implement. Second, it involves only simple operations like matrix-vector multiplications and evaluations of proximal operators, which are for many problems of interest simple and in closed form or easy to compute iteratively; cf., e.g., [33]. However, for large problems that are encountered in many real world applications, even these simple operations might still be too costly to perform very often.

We propose a stochastic extension of PDHG for saddle point problems that are separable in the dual variable (cf., e.g., [18, 53, 55, 34]) where not all, but only a few, of these operations are performed in every iteration. Moreover, as in incremental optimization algorithms [48, 31, 10, 9, 8, 46, 20] over the course of the iterations we continuously build up information from previous iterations, which reduces variance and thereby negative effects of stochasticity. Nonuniform samplings [40, 38, 53, 39, 2] have been proven very efficient for stochastic optimization. In this work we use the expected separable overapproximation framework of [38, 39, 41] to prove all statements for all nontrivial and iteration-independent samplings.

Related work. The proposed algorithm can be seen as a generalization of the algorithm of [18, 55, 53] to arbitrary blocks and a much wider class of samplings. Moreover, in contrast to their results, our results generalize the deterministic case considered in [37, 13, 36, 15]. Fercoq and Bianchi [23] proposed a stochastic primal-dual algorithm with explicit gradient steps that allows for larger step sizes by averaging over previous iterates; however, this comes at the cost of prohibitively large memory requirements. Similar memory issues are encountered by a primal-dual algorithm of [3]. It is related to forward-backward splitting [30] and averaged gradient descent [10, 19] and therefore suffers the same memory issues as the averaged gradient descent. Moreover, Valkonen proposed a stochastic primal-dual algorithm that can exploit partial strong convexity of the saddle point functional [49]. Randomized versions of the alternating direction method of multipliers are discussed, for instance, in [54, 25]. In contrast to other works on stochastic primal-dual algorithms [35, 52], our analysis is not based on Fejér monotonicity [16]. We therefore do not prove almost sure convergence of the sequence but prove a variety of convergence rates depending on strong convexity assumptions instead.

As a word of warning, our contribution should not be mistaken by other "stochastic" primal-dual algorithms, where errors in the computation of matrix-vector products and evaluation of proximal operators are modeled by random variables; cf., e.g., [35, 16, 45]. In our work we deliberately choose to compute only a subset of a whole iteration to save computational cost. These two notations are related but are certainly not the same.

1.1. Contributions.

We briefly mention the main contributions of our work.

Generalization of deterministic case. The proposed stochastic algorithm is a direct generalization of the deterministic setting [37, 13, 36, 14, 15]. In the degenerate case where in every iteration all computations are performed, our algorithm coincides with the original deterministic algorithm. Moreover, the same holds true for our analysis of the stochastic algorithm where we recover almost all deterministic statements [13, 36] in this degenerate case. Therefore, the theorems for both the deterministic and the stochastic cases can be addressed by a single proof.

Better rates. Our analysis extends the simple setting of [53] such that the strong convexity assumptions and the sampling do not have to be uniform. Even in the special case of uniform strong convexity and uniform sampling, the proven

convergence rates are slightly better than the ones proven in [53].

Arbitrary sampling. The proposed algorithm is guaranteed to converge under a very general class of samplings [38, 39, 41] and thereby generalizes also the algorithm of [53], which has only been analyzed for two specific samplings. As long as the sampling is independent and identically distributed (i.i.d.) over the iterations and all computations have nonzero probability to be carried out, the theory holds and the algorithm will converge with the proven convergence rates.

Acceleration. We propose an acceleration of the stochastic primal-dual algorithm which accelerates the convergence from $\mathcal{O}(1/K)$ to $\mathcal{O}(1/K^2)$ if parts of the saddle point functional are strongly convex, thereby resulting in a significantly faster algorithm.

Scaling invariance. In the strongly convex case, we propose parameters for several serial samplings (uniform, importance, optimal), all based on the condition numbers of the problem and thereby independent of scaling.

2. General problem. Let $\mathbb{X}, \mathbb{Y}_i, i = 1, \dots, n$, be real Hilbert spaces of any dimension and define the product space $\mathbb{Y} := \prod_{i=1}^n \mathbb{Y}_i$. For $y \in \mathbb{Y}$, we shall write $y = (y_1, y_2, \dots, y_n)$, where $y_i \in \mathbb{Y}_i$. Further, we consider the natural inner product on the product space $\mathbb{Y}$ given by $\langle y, z\rangle = \sum_{i=1}^n \langle y_i, z_i\rangle$, where $y_i, z_i \in \mathbb{Y}_i$. This inner product induces the norm $\|y\|^2 = \sum_{i=1}^n \|y_i\|^2$. Moreover, for simplicity we will consider the space $\mathbb{W} := \mathbb{X} \times \mathbb{Y}$ that combines both primal and dual variables.

Let $\mathbf{A} : \mathbb{X} \to \mathbb{Y}$ be a bounded linear operator. Due to the product space nature of $\mathbb{Y}$, we have $(\mathbf{A}x)_i = \mathbf{A}_i x$, where $\mathbf{A}_i : \mathbb{X} \to \mathbb{Y}_i$ are linear operators. The adjoint of $\mathbf{A}$ is given by $\mathbf{A}^* y = \sum_{i=1}^n \mathbf{A}_i^* y_i$. Moreover, let $f : \mathbb{Y} \to \mathbb{R}_\infty := \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{X} \to \mathbb{R}_\infty$ be convex functions. In particular, we assume that f is separable, i.e., $f(y) = \sum_{i=1}^n f_i(y_i)$.

Given the setup described above, we consider the optimization problem

$$\min_{x \in \mathbb{X}} \left\{ \Phi(x) := \sum_{i=1}^n f_i(\mathbf{A}_i x) + g(x) \right\}. \tag{1}$$

Instead of solving (1) directly, it is often desirable to reformulate the problem as a saddle point problem with the help of the Fenchel conjugate. If f is proper, convex, and lower semicontinuous, then $f(y) = f^{**}(y) = \sup_{z \in \mathbb{Y}} \langle z, y\rangle - f^*(z)$, where $f^* : \mathbb{Y} \to \mathbb{R} \cup \{-\infty, +\infty\}$, $f^*(y) = \sum_{i=1}^n f_i^*(y_i)$ is the Fenchel conjugate of f (and f^{**} its biconjugate defined as the conjugate of the conjugate). Then solving (1) is equivalent to finding the primal part x of a solution to the saddle point problem (called a saddle point)

$$\min_{x \in \mathbb{X}} \sup_{y \in \mathbb{Y}} \left\{ \Psi(x, y) := \sum_{i=1}^n \langle \mathbf{A}_i x, y_i\rangle - f_i^*(y_i) + g(x) \right\}. \tag{2}$$

We will assume that saddle point problem (2) has a solution. For conditions for existence and uniqueness, we refer the reader to [5]. A saddle point $w^\sharp = (x^\sharp, y^\sharp) = (x^\sharp, y_1^\sharp, \dots, y_n^\sharp) \in \mathbb{W}$ satisfies the optimality conditions

$$\mathbf{A}_i x^\sharp \in \partial f_i^*(y_i^\sharp), \quad i = 1, \dots, n, \qquad -\mathbf{A}^* y^\sharp \in \partial g(x^\sharp).$$

An important notion in this work is *strong convexity*. A functional g is called μ_g-convex if $g - \frac{\mu_g}{2}\|\cdot\|^2$ is convex. In general, we assume that g is μ_g-convex, and f_i^* is

μ_i-convex with nonnegative strong convexity parameters $\mu_g, \mu_i \geq 0$. The convergence results in this contribution cover three different cases of regularity: (i) no strong convexity $\mu_g, \mu_i = 0$, (ii) semistrong convexity $\mu_g > 0$ or $\mu_i > 0$, and (iii) full strong convexity $\mu_g, \mu_i > 0$. For notational convenience we make use of the operator $\mathbf{M} := \operatorname{diag}(\mu_1 \mathbf{I}, \ldots, \mu_n \mathbf{I})$.

A very popular algorithm to solve the saddle point problem (2) is the primal-dual hybrid gradient (PDHG) algorithm [37, 21, 13, 36, 14, 15]. It reads (with extrapolation on y)

$$\begin{aligned} x^{(k+1)} &= \operatorname{prox}_g^\tau \left(x^{(k)} - \tau \mathbf{A}^* \overline{y}^{(k)} \right), \\ y^{(k+1)} &= \operatorname{prox}_{f^*}^\sigma \left(y^{(k)} + \sigma \mathbf{A} x^{(k+1)} \right), \\ \overline{y}^{(k+1)} &= y^{(k+1)} + \theta \left(y^{(k+1)} - y^{(k)} \right), \end{aligned}$$

where the *proximal operator (or proximity/resolvent operator)* is defined as

$$\operatorname{prox}_f^\tau (y) := \arg\min_{x \in \mathbb{X}} \left\{ \frac{1}{2} \|x - y\|_{\tau^{-1}}^2 + f(x) \right\}$$

and the weighted norm by $\|x\|_{\tau^{-1}}^2 = \langle \tau^{-1} x, x \rangle$. Its convergence is guaranteed if the step size parameters σ, τ are positive and satisfy $\sigma\tau \|\mathbf{A}\|^2 < 1, \theta = 1$ [13]. Note that the definition of the proximal operator is well-defined for an *operator-valued* step size τ. In the case of a separable function f and with operator-valued step sizes, PDHG takes the form

$$x^{(k+1)} = \operatorname{prox}_g^{\mathbf{T}} \left(x^{(k)} - \mathbf{T}\mathbf{A}^* \overline{y}^{(k)} \right), \tag{3a}$$

$$y_i^{(k+1)} = \operatorname{prox}_{f_i^*}^{\mathbf{S}_i} \left(y_i^{(k)} + \mathbf{S}_i \mathbf{A}_i x^{(k+1)} \right), \qquad i = 1, \ldots, n, \tag{3b}$$

$$\overline{y}^{(k+1)} = y^{(k+1)} + \theta \left(y^{(k+1)} - y^{(k)} \right). \tag{3c}$$

Here the step size parameters $\mathbf{S} = \operatorname{diag}(\mathbf{S}_1, \ldots, \mathbf{S}_n)$ (a block diagonal operator), $\mathbf{S}_1, \ldots, \mathbf{S}_n$, and $\mathbf{T}$ are symmetric and positive definite. The algorithm is guaranteed to converge if $\|\mathbf{S}^{1/2} \mathbf{A} \mathbf{T}^{1/2}\| < 1$ and $\theta = 1$ [36].

3. Algorithm. In this work we extend PDHG to a stochastic setting where in each iteration we update a random subset $\mathbb{S}$ of the dual variables (3b). This subset is sampled in an i.i.d. fashion from a fixed but otherwise *arbitrary* distribution, whence the name "arbitrary sampling." In order to guarantee convergence, it is necessary to assume that the sampling is "proper" [42, 39]. A sampling is proper if for each dual variable i we have $i \in \mathbb{S}$ with a positive probability $p_i > 0$. Examples of proper samplings include the *full sampling* where $\mathbb{S} = \{1, \ldots, n\}$ with probability 1 and serial sampling where $\mathbb{S} = \{i\}$ is chosen with probability p_i. It is important to note that also other samplings are admissible. For instance, for $n = 3$, consider the sampling that selects $\mathbb{S} = \{1, 2\}$ with probability $1/3$ and $\mathbb{S} = \{2, 3\}$ with probability $2/3$. Then the probabilities for the three blocks are $p_1 = 1/3$, $p_2 = 1$, and $p_3 = 2/3$, which makes it a proper sampling. However, if only $\mathbb{S} = \{1, 2\}$ is chosen with probability 1, then this sampling is not proper, as the probability for the third block is zero: $p_3 = 0$.

The algorithm we propose is formalized as Algorithm 1. As in the original PDHG algorithm, the step size parameters $\mathbf{T}, \mathbf{S}_i$ have to be self-adjoint and positive definite operators for the updates to be well-defined. The extrapolation is performed with a

Algorithm 1 Stochastic primal-dual hybrid gradient algorithm (SPDHG).
Input: $x^{(0)}, y^{(0)}$, $\mathbf{S} = \operatorname{diag}(\mathbf{S}_1, \dots, \mathbf{S}_n)$, $\mathbf{T}, \theta$, $\mathbb{S}^{(k)}$, K. **Initialize:** $\overline{y}^{(0)} = y^{(0)}$.

for $k = 0, \dots, K-1$ **do**

$$x^{(k+1)} = \operatorname{prox}_g^{\mathbf{T}}\left(x^{(k)} - \mathbf{T}\mathbf{A}^*\overline{y}^{(k)}\right)$$

Select $\mathbb{S}^{(k+1)} \subset \{1, \dots, n\}$

$$y_i^{(k+1)} = \begin{cases} \operatorname{prox}_{f_i^*}^{\mathbf{S}_i}\left(y_i^{(k)} + \mathbf{S}_i\mathbf{A}_i x^{(k+1)}\right) & \text{if } i \in \mathbb{S}^{(k+1)} \\ y_i^{(k)} & \text{else} \end{cases}$$

$$\overline{y}^{(k+1)} = y^{(k+1)} + \theta\mathbf{Q}\left(y^{(k+1)} - y^{(k)}\right)$$

end for

scalar $\theta > 0$ and an operator $\mathbf{Q} \coloneqq \operatorname{diag}(p_1^{-1}\mathbf{I}, \dots, p_n^{-1}\mathbf{I})$ of probabilities p_i that an index is selected in each iteration.

Remark 1. Both the primal and dual iterates $x^{(k)}$ and $y^{(k)}$ are random variables, but only the dual iterate $y^{(k)}$ depends on the sampling $\mathbb{S}^{(k)}$. However, $x^{(k)}$ depends of course on all previous samplings $\mathbb{S}^{(i)}, i < k$.

Remark 2. Due to the sampling, each iteration requires both $\mathbf{A}_i$ and $\mathbf{A}_i^*$ to be evaluated only for each selected index $i \in \mathbb{S}^{(k+1)}$. To see this, note that

$$\mathbf{A}^*\overline{y}^{(k+1)} = \mathbf{A}^* y^{(k)} + \sum_{i \in \mathbb{S}^{(k+1)}} \left(1 + \frac{\theta}{p_i}\right) \mathbf{A}_i^*\left(y_i^{(k+1)} - y_i^{(k)}\right),$$

where $\mathbf{A}^* y^{(k)}$ can be stored from the previous iteration (requiring the same memory as the primal variable x) and the operators $\mathbf{A}_i^*$ are evaluated only for $i \in \mathbb{S}^{(k+1)}$.

4. General convex case. We first analyze the convergence of Algorithm 1 in the general convex case without making use of any strong convexity or smoothness assumptions. In order to analyze the convergence for the large class of samplings described in the previous section we make use of the *expected separable overapproximation (ESO)* inequality [39].

DEFINITION 4.1 (expected separable overapproximation). *Let $\mathbb{S} \subset \{1, \dots, n\}$ be a random set and $p_i \coloneqq \mathbb{P}(i \in \mathbb{S})$ the probability that an index i is in the random set $\mathbb{S}$. Moreover, let $\mathbf{C}_i : \mathbb{X} \to \mathbb{Y}_i$ be bounded linear operators and define $\mathbf{C} : \mathbb{X} \to \mathbb{Y} = \prod_{i=1}^n \mathbb{Y}_i$ as $(\mathbf{C}x)_i \coloneqq \mathbf{C}_i x$. Note that its adjoint is given by $\mathbf{C}^* z = \sum_{i=1}^n \mathbf{C}_i^* z_i$. We say that $\{v_i\} \subset \mathbb{R}^n$ fulfill the* ESO inequality *if for all $z \in \mathbb{Y}$ it holds that*

$$\mathbb{E}_{\mathbb{S}} \left\| \sum_{i \in \mathbb{S}} \mathbf{C}_i^* z_i \right\|^2 \le \sum_{i=1}^n p_i v_i \|z_i\|^2 . \tag{4}$$

Such parameters $\{v_i\}$ are called ESO parameters *of $\mathbf{C}$ and $\mathbb{S}$.*

Remark 3. Note that for any bounded linear operator $\mathbf{C}$ such parameters always exist but are obviously not unique. For the efficiency of the algorithm it is desirable to find ESO parameters such that (4) is as tight as possible; i.e., we want the ESO parameters $\{v_i\}$ to be small. As we shall see, the ESO parameters influence the choice of the extrapolation parameter θ in the strongly convex case.

The ESO inequality was first proposed by Richtárik and Takáč [42] to study parallel coordinate descent methods in the context of *uniform* samplings, which are

samplings for which $p_i = p_j$ for all i, j. Improved bounds for ESO parameters were obtained in [24] and used in the context of accelerated coordinate descent. Qu and Richtárik [39] performed an in-depth study of ESO parameters. The ESO inequality is also critical in the study of minibatch stochastic gradient descent with [28] or without [47] variance reduction.

Example 1 (full sampling). Let $\mathbb{S} = \{1, \dots, n\}$ with probability 1 such that $p_i = \mathbb{P}(i \in \mathbb{S}) = 1$ and $\mathbf{C}_i = \mathbf{S}_i^{1/2}\mathbf{A}_i\mathbf{T}^{1/2}$. Then some ESO parameters are given by $v_i = \|\mathbf{C}\|^2$. Thus, the deterministic condition on convergence, $\|\mathbf{S}^{1/2}\mathbf{A}\mathbf{T}^{1/2}\| < 1$, implies a bound on some ESO parameters $v_i < p_i$.

Example 2 (serial sampling). Let $\mathbb{S} = \{i\}$ be chosen with probability $p_i > 0$ and $\mathbf{C}_i = \mathbf{S}_i^{1/2}\mathbf{A}_i\mathbf{T}^{1/2}$. Then some ESO parameters are given by $v_i = \|\mathbf{C}_i\|^2$. Note that obviously $\|\mathbf{C}_i\| \le \|\mathbf{C}\|$ such that the ESO parameters for serial sampling are smaller than the ones for full sampling.

We will frequently need to estimate the expected value of inner products, which we will do by means of ESO parameters. Recall that we defined weighted norms as $\|x\|^2_{\mathbf{T}^{-1}} \coloneqq \langle \mathbf{T}^{-1}x, x\rangle$. The proof of this lemma can be found in the appendix.

LEMMA 4.2. *Let $\mathbb{S} \subset \{1, \dots, n\}$ be a random set, and let $y_i^+ = \hat{y}_i$ if $i \in \mathbb{S}$ and y_i otherwise. Moreover, let $\{v_i\}$ be some ESO parameters of $\mathbf{S}^{1/2}\mathbf{A}\mathbf{T}^{1/2}$ and $p_i = \mathbb{P}(i \in \mathbb{S})$. Then for any $x \in \mathbb{X}$ and $c > 0$*

$$2\mathbb{E}_{\mathbb{S}}\langle \mathbf{Q}\mathbf{A}x, y^+ - y\rangle \ge -\mathbb{E}_{\mathbb{S}}\left\{\frac{1}{c}\|x\|^2_{\mathbf{T}^{-1}} + c\max_i \frac{v_i}{p_i}\|y^+ - y\|^2_{\mathbf{Q}\mathbf{S}^{-1}}\right\}.$$

The analysis for the general convex case will use the notation of *Bregman distance*, which is defined for any function $f : \mathbb{X} \to \mathbb{R}_\infty$, $x, y \in \mathbb{X}$, and $q \in \partial f(y)$ in the subdifferential of f at y as

$$D_f^q(x, y) \coloneqq f(x) - f(y) - \langle q, x - y\rangle.$$

Next to Bregman distances, one can measure optimality by the *partial primal-dual gap*. Let $\mathbb{B}_1 \times \mathbb{B}_2 \subset \mathbb{W} = \mathbb{X} \times \mathbb{Y}$; then we define the partial primal-dual gap as

$$G_{\mathbb{B}_1\times\mathbb{B}_2}(x, y) \coloneqq \sup_{\tilde{y}\in\mathbb{B}_2} \Psi(x, \tilde{y}) - \inf_{\tilde{x}\in\mathbb{B}_1} \Psi(\tilde{x}, y).$$

It is convenient to define $\mathbb{B} \coloneqq \mathbb{B}_1 \times \mathbb{B}_2 \subset \mathbb{W}$ and to denote the gap as $G_{\mathbb{B}}(w) \coloneqq G_{\mathbb{B}_1\times\mathbb{B}_2}(x, y)$. Note that if $\mathbb{B}$ contains a saddle point $w^\sharp = (x^\sharp, y^\sharp)$, then we have that

$$G_{\mathbb{B}}(w) \ge \Psi(x, y^\sharp) - \Psi(x^\sharp, y) = D_g^{-\mathbf{A}^*y^\sharp}(x, x^\sharp) + D_{f^*}^{\mathbf{A}x^\sharp}(y, y^\sharp) = D_h^q(w, w^\sharp) \ge 0,$$

where the first equality is obtained by adding a zero, and we used $h(w) \coloneqq g(x) + f^*(y)$ and $q \coloneqq (-\mathbf{A}^*y^\sharp, \mathbf{A}x^\sharp) \in \partial h(w^\sharp)$ for the last equality. The nonnegativity stems from the fact that Bregman distances of convex functionals are nonnegative and h is convex indeed.

We will make frequent use of the following "distance functions":

$$\mathcal{F}_i(y_i|\tilde{x}, \tilde{y}_i) \coloneqq f_i^*(y_i) - f_i^*(\tilde{y}_i) - \langle \mathbf{A}_i\tilde{x}, y_i - \tilde{y}_i\rangle$$

and $\mathcal{F}(y|\tilde{w}) \coloneqq \sum_{i=1}^n \mathcal{F}_i(y_i|\tilde{x}, \tilde{y}_i)$. Note that these are strongly related to Bregman distances; if $w^\sharp$ is a saddle point, then $\mathcal{F}(y|w^\sharp) = D_{f^*}^{\mathbf{A}x^\sharp}(y, y^\sharp)$ is the Bregman distance

of f^* between y and $y^\sharp$. Similarly, we make use of the weighted distance

$$\mathcal{F}^p(y|\tilde{w}) \coloneqq \sum_{i=1}^n \left(\frac{1}{p_i} - 1\right) \mathcal{F}_i(y_i|\tilde{x}, \tilde{y}_i)$$

and the distance for the primal functional $\mathcal{G}(x|\tilde{w}) \coloneqq g(x) - g(\tilde{x}) - \langle -\mathbf{A}^*\tilde{y}, x - \tilde{x}\rangle$. We note that these distances are also related to the partial primal-dual gap, as with $\mathcal{H}(w|\tilde{w}) := \mathcal{G}(x|\tilde{w}) + \mathcal{F}(y|\tilde{w})$ we have

$$G_{\mathbb{B}}(w) = \sup_{\tilde{w}\in\mathbb{B}} \mathcal{H}(w|\tilde{w})\,.$$

THEOREM 4.3. *Let $\theta = 1$ and $\mathbf{T}, \mathbf{S}$ be chosen so that there exist ESO parameters $\{v_i\}$ of $\mathbf{S}^{1/2}\mathbf{A}\mathbf{T}^{1/2}$ with*

$$v_i < p_i, \qquad i = 1, \dots, n\,. \tag{5}$$

Then the Bregman distance between iterates of Algorithm 1 $w^{(k)} = (x^{(k)}, y^{(k)}) \in \mathbb{W}$ and any saddle point $w^\sharp \in \mathbb{W}$ converges to zero almost surely,

$$D_h^q(w^{(k)}, w^\sharp) \to 0 \quad \textit{a.s.} \tag{6}$$

Moreover, the ergodic sequence $w_{(K)} \coloneqq \frac{1}{K}\sum_{k=1}^K w^{(k)}$ converges with rate $1/K$ in an expected partial primal-dual gap sense; i.e., for any set $\mathbb{B} \coloneqq \mathbb{B}_1 \times \mathbb{B}_2 \subset \mathbb{W}$ it holds that

$$\mathbb{E}\, G_{\mathbb{B}}(w_{(K)}) \le \frac{C_{\mathbb{B}}}{K}, \tag{7}$$

where the constant is given by

$$C_{\mathbb{B}} \coloneqq \sup_{x\in\mathbb{B}_1} \frac{1}{2}\|x^{(0)} - x\|^2_{\mathbf{T}^{-1}} + \sup_{y\in\mathbb{B}_2} \frac{1}{2}\|y^{(0)} - y\|^2_{\mathbf{Q}\mathbf{S}^{-1}} + \sup_{w\in\mathbb{B}} \mathcal{F}^p(y^{(0)}|w)\,. \tag{8}$$

The same rate holds for the expected Bregman distance, $\mathbb{E}D_h^q(w_{(K)}, w^\sharp) \le C_{\{w^\sharp\}}/K$.

Remark 4. The meaning of the convergence (6) in Bregman distance depends on the properties of the function h. In the case that h is strictly convex and $\mathbb{W}$ is finite-dimensional, then (6) implies that $\{w^{(k)}\}$ converges in the norm almost surely to $w^\sharp$. In detail, the additional assumptions imply that $D_h^q(\cdot, w^\sharp)$ is coercive ([4, Proposition 2.5, Fact 2.11] based on [44]), and thus $\{w^{(k)}\}$ is bounded and every subsequence has a convergent subsequence, again denoted by $\{w^{(k)}\}$ and its limit by w. By lower semicontinuity it follows that $D_h^q(w, w^\sharp) = 0$ and thus $w = w^\sharp$ by strict convexity

A similar (weak) convergence statement can be made for general (infinite-dimensional) Hilbert spaces. Here we need to assume coercivity of $D_h^q(\cdot, w^\sharp)$ as it does not follow from strict convexity anymore. This is ensured, for instance, if h is superlinear: $h(w)/\|w\| \to \infty$ as $\|w\| \to \infty$.

If h is not strictly convex, then (6) has to be seen in a more generalized sense. For example, if h is an ℓ^1-norm (and thus not strictly convex), then the Bregman distance between $w^{(k)}$ and $w^\sharp$ is zero if and only if they have the same support and sign. Thus, the convergence statement is related to the support and sign of $w^\sharp$. In the extreme case $h \equiv 0$, then $D_h^q(\cdot, w^\sharp) \equiv 0$ and the convergence statement has no meaning.

The proof of this theorem utilizes a standard inequality for which we provide the proof in the appendix for completeness.

LEMMA 4.4. *Consider the deterministic updates*

$$x^{(k+1)} = \operatorname{prox}_g^{\mathbf{T}_{(k)}}\left(x^{(k)} - \mathbf{T}_{(k)}\mathbf{A}^*\overline{y}^{(k)}\right),$$
$$\hat{y}_i^{(k+1)} = \operatorname{prox}_{f_i^*}^{\mathbf{S}^i_{(k)}}\left(y_i^{(k)} + \mathbf{S}^i_{(k)}\mathbf{A}_i x^{(k+1)}\right), \qquad i = 1, \dots, n,$$

with iteration varying step sizes $\mathbf{T}_{(k)}$ *and* $\mathbf{S}_{(k)} = \operatorname{diag}(\mathbf{S}^1_{(k)}, \dots, \mathbf{S}^n_{(k)})$. *Then for any* $(x, y) \in \mathbb{W}$ *it holds that*

$$\begin{aligned}
&\|x^{(k)} - x\|^2_{\mathbf{T}^{-1}_{(k)}} + \|y^{(k)} - y\|^2_{\mathbf{S}^{-1}_{(k)}} \\
&\geq \|x^{(k+1)} - x\|^2_{\mathbf{T}^{-1}_{(k)} + \mu_g \mathbf{I}} + \|\hat{y}^{(k+1)} - y\|^2_{\mathbf{S}^{-1}_{(k)} + \mathbf{M}} \\
&\quad + 2\left(\mathcal{G}(x^{(k+1)}|w) + \mathcal{F}(\hat{y}^{(k+1)}|w)\right) - 2\langle \mathbf{A}(x^{(k+1)} - x), \hat{y}^{(k+1)} - \overline{y}^{(k)}\rangle \\
&\quad + \|x^{(k+1)} - x^{(k)}\|^2_{\mathbf{T}^{-1}_{(k)}} + \|\hat{y}^{(k+1)} - y^{(k)}\|^2_{\mathbf{S}^{-1}_{(k)}}.
\end{aligned}$$

Proof of Theorem 4.3. The result of Lemma 4.4 (with constant step sizes) has to be adapted to the stochastic setting as the dual iterate is updated only with a certain probability. First, a trivial observation is that for any mapping φ it holds that

$$\begin{aligned}
\varphi(\hat{y}_i^{(k+1)}) &= \frac{1}{p_i}\mathbb{E}^{(k+1)}\varphi(y_i^{(k+1)}) - \left(\frac{1}{p_i} - 1\right)\varphi(y_i^{(k)}) \\
&= \left(\frac{1}{p_i} - 1\right)\mathbb{E}^{(k+1)}\varphi(y_i^{(k+1)}) - \left(\frac{1}{p_i} - 1\right)\varphi(y_i^{(k)}) + \mathbb{E}^{(k+1)}\varphi(y_i^{(k+1)}). \qquad (9)
\end{aligned}$$

Thus, for the generalized distance of f^* we arrive at

$$\mathcal{F}(\hat{y}^{(k+1)}|w) = \mathbb{E}^{(k+1)}\mathcal{F}^p(y^{(k+1)}|w) - \mathcal{F}^p(y^{(k)}|w) + \mathbb{E}^{(k+1)}\mathcal{F}(y^{(k+1)}|w), \qquad (10)$$

and for any block diagonal matrix $\mathbf{B} = \operatorname{diag}(\mathbf{B}_1, \dots, \mathbf{B}_n)$

$$\|\hat{y}^{(k+1)} - \cdot\|^2_{\mathbf{B}} = \mathbb{E}^{(k+1)}\|y^{(k+1)} - \cdot\|^2_{\mathbf{QB}} - \|y^{(k)} - \cdot\|^2_{(\mathbf{Q}-\mathbf{I})\mathbf{B}}, \qquad (11)$$

$$\hat{y}^{(k+1)} = \mathbf{Q}\mathbb{E}^{(k+1)}y^{(k+1)} - (\mathbf{Q} - \mathbf{I})y^{(k)}. \qquad (12)$$

Using (10)–(12), we can rewrite the estimate of Lemma 4.4 as

$$\begin{aligned}
&\|x^{(k)} - x\|^2_{\mathbf{T}^{-1}} + \|y^{(k)} - y\|^2_{\mathbf{QS}^{-1}} + 2\mathcal{F}^p(y^{(k)}|w) \\
&\geq \mathbb{E}^{(k+1)}\Big\{\|x^{(k+1)} - x\|^2_{\mathbf{T}^{-1}} + \|y^{(k+1)} - y\|^2_{\mathbf{QS}^{-1}} + 2\mathcal{F}^p(y^{(k+1)}|w) \\
&\quad + 2\mathcal{H}(w^{(k+1)}|w) - 2\langle \mathbf{A}(x^{(k+1)} - x), \mathbf{Q}(y^{(k+1)} - y^{(k)}) + y^{(k)} - \overline{y}^{(k)}\rangle \\
&\quad + \|x^{(k+1)} - x^{(k)}\|^2_{\mathbf{T}^{-1}} + \|y^{(k+1)} - y^{(k)}\|^2_{\mathbf{QS}^{-1}}\Big\}, \qquad (13)
\end{aligned}$$

where we have used the identity

$$\|\cdot\|^2_{\mathbf{B}} + \|\cdot\|^2_{\mathbf{D}} = \|\cdot\|^2_{\mathbf{B}+\mathbf{D}} \qquad (14)$$

to simplify the expression. With the extrapolation $\overline{y}^{(k)} = y^{(k)} + \mathbf{Q}(y^{(k)} - y^{(k-1)})$, the inner product term can be reformulated as

$$\begin{aligned}
&- \langle \mathbf{A}(x^{(k+1)} - x), \mathbf{Q}(y^{(k+1)} - y^{(k)}) + y^{(k)} - \overline{y}^{(k)} \rangle \\
&= -\langle \mathbf{QA}(x^{(k+1)} - x), y^{(k+1)} - y^{(k)} \rangle + \langle \mathbf{QA}(x^{(k+1)} - x), y^{(k)} - y^{(k-1)} \rangle \\
&= -\langle \mathbf{QA}(x^{(k+1)} - x), y^{(k+1)} - y^{(k)} \rangle + \langle \mathbf{QA}(x^{(k)} - x), y^{(k)} - y^{(k-1)} \rangle \\
&\quad + \langle \mathbf{QA}(x^{(k+1)} - x^{(k)}), y^{(k)} - y^{(k-1)} \rangle, \qquad (15)
\end{aligned}$$

and with Lemma 4.2 and $\gamma^2 \coloneqq \max_i v_i/p_i$ it holds that

$$\begin{aligned}
&2\mathbb{E}^{(k)} \langle \mathbf{QA}(x^{(k+1)} - x^{(k)}), y^{(k)} - y^{(k-1)} \rangle \\
&\geq -\mathbb{E}^{(k)} \left\{ \gamma^2 \|x^{(k+1)} - x^{(k)}\|^2_{\mathbf{T}^{-1}} + \|y^{(k)} - y^{(k-1)}\|^2_{\mathbf{QS}^{-1}} \right\}. \qquad (16)
\end{aligned}$$

Taking expectations with respect to $\mathbb{S}^1, \dots, \mathbb{S}^K$ (denoted by $\mathbb{E}$) on (13), using the estimates (15) and (16), and denoting

$$\begin{aligned}
\Delta^{(k)} \coloneqq \mathbb{E} \Big\{ &\|x^{(k)} - x\|^2_{\mathbf{T}^{-1}} + \|y^{(k)} - y\|^2_{\mathbf{QS}^{-1}} + 2\mathcal{F}^p(y^{(k)}|w) \\
&+ \|y^{(k)} - y^{(k-1)}\|^2_{\mathbf{QS}^{-1}} - 2\langle \mathbf{QA}(x^{(k)} - x), y^{(k)} - y^{(k-1)} \rangle \Big\}
\end{aligned}$$

leads with $\gamma < 1$ (follows directly from (5)) to

$$\begin{aligned}
\Delta^{(k)} &\geq \Delta^{(k+1)} + \mathbb{E} \left(2\mathcal{H}(w^{(k+1)}|w) + (1 - \gamma^2)\|x^{(k+1)} - x^{(k)}\|^2_{\mathbf{T}^{-1}} \right) \\
&\geq \Delta^{(k+1)} + 2\mathbb{E}\mathcal{H}(w^{(k+1)}|w). \qquad (17)
\end{aligned}$$

Summing (17) over $k = 0, \dots, K-1$ (note that $y^{(-1)} = y^{(0)}$) and using the estimate (which follows directly from Lemma 4.2)

$$\begin{aligned}
\Delta^{(K)} &\geq \mathbb{E} \left\{ (1 - \gamma^2)\|x^{(K)} - x\|^2_{\mathbf{T}^{-1}} + \|y^{(K)} - y\|^2_{\mathbf{QS}^{-1}} + 2\mathcal{F}^p(y^{(K)}|w) \right\} \\
&\geq 2\mathbb{E}\mathcal{F}^p(y^{(K)}|w)
\end{aligned}$$

yields

$$\mathbb{E} \left\{ \mathcal{F}^p(y^{(K)}|w) + \sum_{k=1}^{K} \mathcal{H}(w^{(k)}|w) \right\} \leq \frac{\Delta^{(0)}}{2}. \qquad (18)$$

All assertions of the theorem follow from inequality (18). Inserting a saddle point $w = w^\sharp$ and taking the limit $K \to \infty$, it follows from (18) that $\mathbb{E} \sum_{k=1}^{\infty} D_h^q(w^{(k)}, w^\sharp) < \infty$, which implies almost surely $\sum_{k=1}^{\infty} D_h^q(w^{(k)}, w^\sharp) < \infty$ and thus (6).

To see (7), note first that

$$\mathcal{F}^p(y^{(0)}|w) - \mathcal{F}^p(y^{(K)}|w) = \mathcal{F}^p(y^{(0)}|x, y^{(K)}) \leq \sup_{w \in \mathbb{B}} \mathcal{F}^p(y^{(0)}|w)$$

and $\Delta^{(0)}/2 - \mathcal{F}^p(y^{(K)}|w) \leq C_{\mathbb{B}}$ if $w \in \mathbb{B}$ with $C_{\mathbb{B}}$ as defined in (8). Moreover, the generalized distance $\mathcal{H}(\cdot|w)$ is convex, and thus dividing (18) by K yields

$$\mathbb{E}\mathcal{H}(w_{(K)}|w) \leq \frac{1}{K}\mathbb{E} \sum_{k=1}^{K} \mathcal{H}(w^{(k)}|w) \leq \frac{C_{\mathbb{B}}}{K}$$

for any $w \in \mathbb{B}$. Taking the supremum over $w \in \mathbb{B}$ yields (7). Noting that $D_h^q(w, w^\sharp) = G_{\{w^\sharp\}}(w)$ completes the proof. ∎

5. Semistrongly convex case. In this section we propose randomized and accelerated algorithms which can exploit strong convexity in either f_i^* or g. Algorithm 2 converges in the dual variable with rate $\mathcal{O}(1/K^2)$ if the convex conjugate f_i^* is strongly convex. Similarly, Algorithm 3 converges with the same accelerated rate $\mathcal{O}(1/K^2)$ in the primal variable if g is strongly convex. For simplicity we restrict ourselves from now on to scalar-valued step sizes, i.e., $\mathbf{T} = \tau\mathbf{I}$ and $\mathbf{S}_i = \sigma_i\mathbf{I}$. However, large parts of what follows hold true for operator-valued step sizes, too.

Algorithm 2 Stochastic primal-dual hybrid gradient algorithm with acceleration on the dual variable (DA-SPDHG).
Input: $x^{(0)}, y^{(0)}$, $\tau_{(0)} \in \mathbb{R}, \tilde{\sigma}_{(0)} \in \mathbb{R}$, $\mathbb{S}^{(k)}$, K. **Initialize:** $\overline{y}^{(0)} = y^{(0)}$.

1: **for** $k = 0, \dots, K-1$ **do**
2: $\quad x^{(k+1)} = \operatorname{prox}_g^{\tau_{(k)}}\left(x^k - \tau_{(k)}\mathbf{A}^*\overline{y}^{(k)}\right)$
3: $\quad$ Select $\mathbb{S}^{(k+1)} \subset \{1, \dots, n\}$
4: $\quad \sigma_i^{(k)} = \frac{\tilde{\sigma}_{(k)}}{\mu_i[p_i - 2(1-p_i)\tilde{\sigma}_{(k)}]}, \quad i \in \mathbb{S}^{(k+1)}$
5: $\quad y_i^{(k+1)} = \begin{cases} \operatorname{prox}_{f_i^*}^{\sigma_i^{(k)}}\left(y_i^{(k)} + \sigma_i^{(k)}\mathbf{A}_i x^{(k+1)}\right) & \text{if } i \in \mathbb{S}^{(k+1)} \\ y_i^{(k)} & \text{else} \end{cases}$
6: $\quad \theta_{(k)} = (1 + 2\tilde{\sigma}_{(k)})^{-1/2}, \quad \tau_{(k+1)} = \tau_{(k)}/\theta_{(k)}, \quad \tilde{\sigma}_{(k+1)} = \theta_{(k)}\tilde{\sigma}_{(k)}$
7: $\quad \overline{y}^{(k+1)} = y^{(k+1)} + \theta_{(k)}\mathbf{Q}\left(y^{(k+1)} - y^{(k)}\right)$
8: **end for**

THEOREM 5.1 (dual strong convexity). *Let f_i^* be strongly convex with constants $\mu_i > 0$, $i = 1, \dots, n$. Consider Algorithm 2 and let the initial step sizes $\tilde{\sigma}_{(0)}, \tau_{(0)}$ be chosen such that*

$$\tilde{\sigma}_{(0)} < \min_i \frac{p_i}{2(1-p_i)}, \tag{19}$$

and for the ESO parameters $\{v_i\}$ of $\mathbf{S}_{(0)}^{1/2}\mathbf{A}\tau_{(0)}^{1/2}$ it holds that

$$v_i \le p_i, \qquad i = 1, \dots, n, \tag{20}$$

with $[\mathbf{S}_{(k)}]_i = \sigma_i^{(k)}\mathbf{I}$ and

$$\sigma_i^{(k)} = \frac{\tilde{\sigma}_{(k)}}{\mu_i[p_i - 2(1-p_i)\tilde{\sigma}_{(k)}]}. \tag{21}$$

Then there exists $\tilde{K} \in \mathbb{N}$ such that for all $K \ge \tilde{K}$ it holds that

$$\mathbb{E}\|y^{(K)} - y^\sharp\|_{\mathbf{Y}_{(0)}}^2 \le \frac{2}{K^2}\left(\|x^{(0)} - x^\sharp\|_{\tau_{(0)}^{-1}}^2 + \|y^{(0)} - y^\sharp\|_{\mathbf{Y}_{(0)}}^2\right),$$

where the metric on $\mathbb{Y}$ is defined by $\mathbf{Y}_{(k)} \coloneqq \mathbf{Q}\mathbf{S}_{(k)}^{-1} + 2\mathbf{M}(\mathbf{Q} - \mathbf{I})$.

Remark 5. As already noted in [13], $\tilde{K}$ is usually fairly small so that the estimate in Theorem 5.1 has practical relevance.

Remark 6. For serial sampling the condition on the ESO parameters (20) is equivalent to

$$\tilde{\sigma}_{(0)} \le \min_i \frac{\mu_i p_i^2}{\tau_{(0)}\|\mathbf{A}_i\|^2 + 2\mu_i p_i(1-p_i)}.$$

In particular, it implies condition (19) on $\tilde{\sigma}_{(0)}$.

This theorem requires an estimate on the expected contraction similar to the proof of Theorem 4.3 and shown in the appendix.

LEMMA 5.2. *Let $x^{(k+1)}, \hat{y}^{(k+1)}$ be defined as in Lemma 4.4, and let $y_i^{(k+1)} = \hat{y}_i^{(k+1)}$ with probability $p_i > 0$ and unchanged otherwise. Moreover, let*

$$\overline{y}^{(k+1)} = y^{(k+1)} + \theta_{(k)}\mathbf{Q}\left(y^{(k+1)} - y^{(k)}\right) \tag{22}$$

and $\{v_i\}$ be some ESO parameters of $\mathbf{S}_{(k)}^{1/2}\mathbf{A}\tau_{(k)}^{1/2}$. Then with $\gamma^2 = \max_i \frac{v_i}{p_i}$ it holds that

$$\begin{aligned}
&\mathbb{E}^{(k,k-1)}\Big\{\|x^{(k)} - x^\sharp\|^2_{\tau_{(k)}^{-1}} + \|y^{(k)} - y^\sharp\|^2_{\mathbf{Q}\mathbf{S}_{(k)}^{-1}+2\mathbf{M}(\mathbf{Q}-\mathbf{I})} \\
&\quad - 2\theta_{(k-1)}\langle\mathbf{Q}\mathbf{A}(x^{(k)} - x^\sharp), y^{(k)} - y^{(k-1)}\rangle + (\gamma\theta_{(k-1)})^2\|y^{(k)} - y^{(k-1)}\|^2_{\mathbf{Q}\mathbf{S}_{(k)}^{-1}}\Big\} \\
&\geq \mathbb{E}^{(k+1,k)}\Big\{\|x^{(k+1)} - x^\sharp\|^2_{\tau_{(k)}^{-1}+2\mu_g\mathbf{I}} + \|y^{(k+1)} - y^\sharp\|^2_{\mathbf{Q}\mathbf{S}_{(k)}^{-1}+2\mathbf{M}\mathbf{Q}} \\
&\quad - 2\langle\mathbf{Q}\mathbf{A}(x^{(k+1)} - x^\sharp), y^{(k+1)} - y^{(k)}\rangle + \|y^{(k+1)} - y^{(k)}\|^2_{\mathbf{Q}\mathbf{S}_{(k)}^{-1}}\Big\}.
\end{aligned}$$

Proof of Theorem 5.1. The update on the step sizes in Algorithm 2 implies that

$$\begin{aligned}
\theta_{(k)}\frac{1}{\tau_{(k)}} &\geq \frac{1}{\tau_{(k+1)}}, \\
\theta_{(k)}\left(\frac{1}{p_i\sigma_i^{(k)}} + \frac{2\mu_i}{p_i}\right) &\geq \frac{1}{p_i\sigma_i^{(k+1)}} + \frac{2(1-p_i)\mu_i}{p_i} \qquad (23)
\end{aligned}$$

for all $i = 1, \ldots, n$ and therefore

$$\theta_{(k)}\|\cdot\|^2_{\tau_{(k)}^{-1}} \geq \|\cdot\|^2_{\tau_{(k+1)}^{-1}}, \tag{24}$$

$$\theta_{(k)}\|\cdot\|^2_{\mathbf{Q}\mathbf{S}_{(k)}^{-1}+2\mathbf{M}\mathbf{Q}} \geq \|\cdot\|^2_{\mathbf{Q}\mathbf{S}_{(k+1)}^{-1}+2\mathbf{M}(\mathbf{Q}-I)} = \|\cdot\|^2_{\mathbf{Y}_{(k+1)}}. \tag{25}$$

To see (23), the auxiliary sequence $\tilde{\sigma}_{(k)}$ satisfies

$$\tilde{\sigma}_{(k)} = \frac{p_i\mu_i\sigma_i^{(k)}}{1 + 2(1-p_i)\mu_i\sigma_i^{(k)}}$$

such that (23) is satisfied as soon as

$$\theta_{(k)}\frac{1 + 2\tilde{\sigma}_{(k)}}{\tilde{\sigma}_{(k)}} \geq \frac{1}{\tilde{\sigma}_{(k+1)}}. \tag{26}$$

Note that the transformation from $\tilde{\sigma}_{(k)}$ to $\sigma_i^{(k)}$ is well-defined if $\tilde{\sigma}_{(k)} < \min_i \frac{p_i}{2(1-p_i)}$, which is the case as $\tilde{\sigma}_{(k)}$ is monotonically nonincreasing and $\tilde{\sigma}_{(0)}$ satisfies the condition. By construction of the sequence $\tilde{\sigma}_{(k+1)} = \theta_{(k)}\tilde{\sigma}_{(k)}$, (26) is solved with equality by $\theta_{(k)} = (1 + 2\tilde{\sigma}_{(k)})^{-1/2}$. Moreover, the sequence $\sigma_i^{(k)}$ is also nonincreasing as

$$\sigma_i^{(k+1)} = \frac{\theta_{(k)}\sigma_i^{(k)}}{1 + 2(1-\theta_{(k)})(1-p_i)\mu_i\sigma_i^{(k)}} \leq \theta_{(k)}\sigma_i^{(k)};$$

thus, with (20) we see that the ESO parameters of $\mathbf{S}_{(k)}^{1/2}\mathbf{A}\tau_{(k)}^{1/2}$ are also bounded by p_i.

For the actual proof of the theorem, note that inequalities (24) and (25) imply

$$\theta_{(k)}\mathbb{E}\Big\{\|x^{(k+1)}-x^\sharp\|^2_{\tau_{(k)}^{-1}}+\|y^{(k+1)}-y^\sharp\|^2_{\mathbf{Q}\mathbf{S}_{(k)}^{-1}+2\mathbf{M}\mathbf{Q}} - 2\langle\mathbf{Q}\mathbf{A}(x^{(k+1)}-x^\sharp),y^{(k+1)}-y^{(k)}\rangle\Big\}\geq\mathbb{E}\Delta^{(k+1)} \tag{27}$$

with

$$\Delta^{(k)}:=\|x^{(k)}-x^\sharp\|^2_{\tau_{(k)}^{-1}}+\|y^{(k)}-y^\sharp\|^2_{\mathbf{Y}_{(k)}}-2\theta_{(k-1)}\langle\mathbf{Q}\mathbf{A}(x^{(k)}-x^\sharp),y^{(k)}-y^{(k-1)}\rangle\,.$$

Thus, combining Lemma 5.2 ($\mu_g=0$) and (27) yields

$$\begin{aligned}
&\theta_{(k)}\mathbb{E}\left\{\Delta^{(k)}+(\gamma\theta_{(k-1)})^2\|y^{(k)}-y^{(k-1)}\|^2_{\mathbf{Q}\mathbf{S}_{(k)}^{-1}}\right\}\\
&\geq\theta_{(k)}\mathbb{E}\Big\{\|x^{(k+1)}-x^\sharp\|^2_{\tau_{(k)}^{-1}}+\|y^{(k+1)}-y^\sharp\|^2_{\mathbf{Q}\mathbf{S}_{(k)}^{-1}+2\mathbf{M}\mathbf{Q}}\\
&\qquad-2\langle\mathbf{Q}\mathbf{A}(x^{(k+1)}-x^\sharp),y^{(k+1)}-y^{(k)}\rangle+\|y^{(k+1)}-y^{(k)}\|^2_{\mathbf{Q}\mathbf{S}_{(k)}^{-1}}\Big\}\\
&\geq\mathbb{E}\left\{\Delta^{(k+1)}+\theta_{(k)}\|y^{(k+1)}-y^{(k)}\|^2_{\mathbf{Q}\mathbf{S}_{(k)}^{-1}}\right\}.
\end{aligned}$$

With $\gamma\theta_{(k-1)}\leq 1,\mathbf{S}_{(k+1)}\leq\theta_{(k)}\mathbf{S}_{(k)}$, and $\bar\Delta^{(k)}:=\mathbb{E}\{\Delta^{(k)}+\|y^{(k)}-y^{(k-1)}\|^2_{\mathbf{Q}\mathbf{S}_{(k)}^{-1}}\}$ we derive the recursion

$$\begin{aligned}
\theta_{(k)}\bar\Delta^{(k)}&\geq\theta_{(k)}\mathbb{E}\left\{\Delta^{(k)}+(\gamma\theta_{(k-1)})^2\|y^{(k)}-y^{(k-1)}\|^2_{\mathbf{Q}\mathbf{S}_{(k)}^{-1}}\right\}\\
&\geq\mathbb{E}\left\{\Delta^{(k+1)}+\theta_{(k)}\|y^{(k+1)}-y^{(k)}\|^2_{\mathbf{Q}\mathbf{S}_{(k)}^{-1}}\right\}\geq\bar\Delta^{(k+1)}.
\end{aligned}$$

Using this inequality recursively, $y^{(-1)}=y^{(0)}$, we arrive at

$$\begin{aligned}
\prod_{k=0}^{K-1}\theta_{(k)}\bar\Delta^{(0)}\geq\bar\Delta^{(K)}&\geq\mathbb{E}\left\{(1-\gamma^2)\|x^{(K)}-x^\sharp\|^2_{\tau_{(K)}^{-1}}+\|y^{(K)}-y^\sharp\|^2_{\mathbf{Y}_{(K)}}\right\}\\
&\geq\mathbb{E}\|y^{(K)}-y^\sharp\|^2_{\mathbf{Y}_{(K)}},
\end{aligned}$$

where the second inequality follows directly from Lemma 4.2 and the third inequality from $\gamma\leq 1$, which holds by assumption (20).

As $\bar\Delta^{(0)}=\|x^{(0)}-x^\sharp\|^2_{\tau_{(0)}^{-1}}+\|y^{(0)}-y^\sharp\|^2_{\mathbf{Y}_{(0)}}$, $\theta_{(k)}=\frac{\tilde\sigma_{(k+1)}}{\tilde\sigma_{(k)}}$, and

$$\|\cdot\|^2_{\mathbf{Y}_{(K)}}=\frac{1}{\tilde\sigma_{(K)}}\|\cdot\|^2_{\mathbf{M}}=\frac{\tilde\sigma_{(0)}}{\tilde\sigma_{(K)}}\|\cdot\|^2_{\mathbf{Y}_{(0)}},$$

which holds by the definition of $\tilde\sigma_{(k)}$, it holds that

$$\mathbb{E}\|y^{(K)}-y^\sharp\|^2_{\mathbf{Y}_{(0)}}\leq\left(\frac{\tilde\sigma_{(K)}}{\tilde\sigma_{(0)}}\right)^2\left\{\|x^{(0)}-x^\sharp\|^2_{\tau_{(0)}^{-1}}+\|y^{(0)}-y^\sharp\|^2_{\mathbf{Y}_{(0)}}\right\}.$$

Finally, the assertion follows by Corollary 1 of [13]. ∎

Remark 7. If g is strongly convex, then the primal variable can be accelerated; see Algorithm 3. Its convergence can be analyzed similarly to the deterministic case (cf. Appendix C.2 of [14]) and omitted is here for brevity. It converges with rate $\mathcal{O}(1/K^2)$ in the primal variable if the ESO parameters satisfy $v_i < p_i$.

Algorithm 3 Stochastic primal-dual hybrid gradient algorithm with acceleration on the primal variable (PA-SPDHG).
Input: $x^{(0)}, y^{(0)}, \tau_{(0)} \in \mathbb{R}, \sigma^{(0)} \in \mathbb{R}^n$, $\mathbb{S}^{(k)}$, K. **Initialize:** $\overline{y}^{(0)} = y^{(0)}$.

1: **for** $k = 0, \dots, K-1$ **do**
2: $\quad x^{(k+1)} = \operatorname{prox}_g^{\tau_{(k)}}\left(x^{(k)} - \tau_{(k)}\mathbf{A}^*\overline{y}^{(k)}\right)$
3: $\quad$ Select $\mathbb{S}^{(k+1)} \subset \{1, \dots, n\}$
4: $\quad y_i^{(k+1)} = \begin{cases} \operatorname{prox}_{f_i^*}^{\sigma_i^{(k)}}\left(y_i^{(k)} + \sigma_i^{(k)}\mathbf{A}_i x^{(k+1)}\right) & \text{if } i \in \mathbb{S}^{(k+1)} \\ y_i^{(k)} & \text{else} \end{cases}$
5: $\quad \theta_{(k)} = (1 + 2\mu_g\tau_{(k)})^{-1/2}, \quad \tau_{(k+1)} = \theta_{(k)}\tau_{(k)}, \quad \sigma^{(k+1)} = \sigma^{(k)}/\theta_{(k)}$
6: $\quad \overline{y}^{(k+1)} = y^{(k+1)} + \theta_{(k)}\mathbf{Q}\left(y^{(k+1)} - y^{(k)}\right)$
7: **end for**

6. Strongly convex case. If both f_i^* and g are strongly convex, we may find step size parameters such that Algorithm 1 converges linearly.

THEOREM 6.1. *Let $(x^\sharp, y^\sharp) \in \mathbb{W}$ be a saddle point and g, f_i^* be strongly convex with constants $\mu_g, \mu_i > 0, i = 1, \dots, n$. Let the step sizes $\tau, \sigma_1, \dots, \sigma_n, 0 < \theta < 1$ be chosen such that the ESO parameters $\{v_i\}$ of $\mathbf{S}^{1/2}\mathbf{A}\tau^{1/2}$ can be estimated as*

$$v_i < \frac{p_i}{\theta}, \qquad i = 1, \dots, n, \tag{28}$$

and the extrapolation θ satisfies the lower bounds

$$\theta \geq \frac{1}{1 + 2\mu_g\tau}, \qquad \theta \geq \frac{1 + 2(1 - p_i)\mu_i\sigma_i}{1 + 2\mu_i\sigma_i}, \qquad i = 1, \dots, n. \tag{29}$$

Then the iterates of Algorithm 1 *converge linearly to the saddle point; in particular*

$$\mathbb{E}\left\{(1 - \gamma^2\theta)\|x^{(K)} - x^\sharp\|_{\mathbf{X}}^2 + \|y^{(K)} - y^\sharp\|_{\mathbf{Y}}^2\right\} \leq \theta^K\left\{\|x^{(0)} - x^\sharp\|_{\mathbf{X}}^2 + \|y^{(0)} - y^\sharp\|_{\mathbf{Y}}^2\right\}$$

holds where the metrics are given by $\mathbf{X} := (\tau^{-1} + 2\mu_g)\mathbf{I}$, $\mathbf{Y} := (\mathbf{S}^{-1} + 2\mathbf{M})\mathbf{Q}$, and $\gamma^2 = \max_i v_i/p_i$.

Proof. The requirements (29) on the step sizes $\tau, \sigma_1, \dots, \sigma_n$ and θ imply $\theta\|\cdot\|_{\mathbf{X}}^2 \geq \|\cdot\|_{\tau^{-1}}^2$ and $\theta\|\cdot\|_{\mathbf{Y}}^2 \geq \|\cdot\|_{\mathbf{Q}\mathbf{S}^{-1} + 2\mathbf{M}(\mathbf{Q}-\mathbf{I})}^2$. Thus, we directly get

$$\begin{aligned}\theta\,\mathbb{E}\Delta^{(k)} \geq \mathbb{E}\Big\{&\|x^{(k)} - x^\sharp\|_{\tau^{-1}}^2 + \|y^{(k)} - y^\sharp\|_{\mathbf{Q}\mathbf{S}^{-1} + 2\mathbf{M}(\mathbf{Q}-\mathbf{I})}^2 \\ &- 2\theta\langle\mathbf{Q}\mathbf{A}(x^{(k)} - x^\sharp), y^{(k)} - y^{(k-1)}\rangle\Big\},\end{aligned} \tag{30}$$

where we denoted

$$\Delta^{(k)} := \|x^{(k)} - x^\sharp\|_{\mathbf{X}}^2 + \|y^{(k)} - y^\sharp\|_{\mathbf{Y}}^2 - 2\langle\mathbf{Q}\mathbf{A}(x^{(k)} - x^\sharp), y^{(k)} - y^{(k-1)}\rangle\,.$$

Combining (30) and Lemma 5.2 with constant step sizes yields

$$\theta\,\mathbb{E}\Delta^{(k)} \geq \mathbb{E}\left\{\Delta^{(k+1)} + \|y^{(k+1)} - y^{(k)}\|^2_{\mathbf{Q}\mathbf{S}^{-1}} - (\gamma\theta)^2\|y^{(k)} - y^{(k-1)}\|^2_{\mathbf{Q}\mathbf{S}^{-1}}\right\}.$$

Multiplying both sides by $\theta^{-(k+1)}$ and summing over $k = 0, \dots, K-1$ yields

$$\begin{aligned}
\Delta^{(0)} &\geq \theta^{-K}\mathbb{E}\Big\{\Delta^{(K)} + \|y^{(K)} - y^{(K-1)}\|^2_{\mathbf{Q}\mathbf{S}^{-1}}\Big\} \\
&\quad + (1-\gamma^2\theta)\mathbb{E}\sum_{k=1}^{K-1}\theta^{-k}\|y^{(k)} - y^{(k-1)}\|^2_{\mathbf{Q}\mathbf{S}^{-1}} \\
&\geq \theta^{-K}\mathbb{E}\Big\{\|x^{(K)} - x^\sharp\|^2_{\mathbf{X}} + \|y^{(K)} - y^\sharp\|^2_{\mathbf{Y}} + \|y^{(K)} - y^{(K-1)}\|^2_{\mathbf{Q}\mathbf{S}^{-1}} \\
&\qquad - 2\langle\mathbf{Q}\mathbf{A}(x^{(K)} - x^\sharp), y^{(K)} - y^{(K-1)}\rangle\Big\} \\
&\geq \theta^{-K}\mathbb{E}\left\{\|x^{(K)} - x^\sharp\|^2_{\mathbf{X}} - \gamma^2\|x^{(K)} - x^\sharp\|^2_{\tau^{-1}} + \|y^{(K)} - y^\sharp\|^2_{\mathbf{Y}}\right\} \\
&\geq \theta^{-K}\mathbb{E}\left\{(1-\gamma^2\theta)\|x^{(K)} - x^\sharp\|^2_{\mathbf{X}} + \|y^{(K)} - y^\sharp\|^2_{\mathbf{Y}}\right\},
\end{aligned}$$

where we used again Lemma 4.2 and the nonnegativity of norms for the second inequality. Thus, the assertion is proved. ∎

6.1. Optimal parameters for serial sampling. This analysis is to optimize the convergence rate θ of Theorem 6.1 for three different serial sampling options where exactly one block is chosen in each iteration. Other sampling strategies, including multiblock, parallel, etc. [39], will be subject of future work.

We will derive the rates and parameters in terms of the *condition numbers* $\kappa_i := \|\mathbf{A}_i\|^2/(\mu_g\mu_i)$ as these are scaling invariant, and thus we cannot improve the rates by simple rescaling of the problem. This can be seen as follows. If we rewrite problem (2) in terms of the scaled variables $\overline{x} := \alpha x$ and $\overline{y}_i := \beta_i y_i$, then the corresponding operators $\overline{\mathbf{A}}_i := \mathbf{A}_i/(\alpha\beta_i)$ have norm $\|\overline{\mathbf{A}}_i\| = \|\mathbf{A}_i\|/(\alpha\beta_i)$, the function $\overline{g}(\overline{x}) := g(\overline{x}/\alpha)$ is $\overline{\mu}_g := \mu_g/\alpha^2$ strongly convex, and the functions $\overline{f}_i^*(\overline{y}_i) := f_i^*(\overline{y}_i/\beta_i)$ are $\overline{\mu}_i := \mu_i/\beta_i^2$ strongly convex. Thus the condition numbers are scaling invariant as

$$\overline{\kappa}_i = \frac{\|\overline{\mathbf{A}}_i\|^2}{\overline{\mu}_g\overline{\mu}_i} = \frac{\frac{1}{(\alpha\beta_i)^2}\|\mathbf{A}_i\|^2}{\frac{1}{\alpha^2}\mu_g\frac{1}{\beta_i^2}\mu_i} = \frac{\|\mathbf{A}_i\|^2}{\mu_g\mu_i} = \kappa_i\,.$$

With $\bar{\sigma}_i := \sigma_i\mu_i$ and $\bar{\tau} := \tau\mu_g$, the conditions on the step sizes (29) become

$$\theta \geq \frac{1}{1+2\bar{\tau}}\,, \quad \theta \geq \max_i 1 - 2\frac{\bar{\sigma}_i p_i}{1+2\bar{\sigma}_i}\,, \quad \text{and} \quad \max_i \bar{\tau}\bar{\sigma}_i\kappa_i\theta \leq \rho^2 p_i \tag{31}$$

for some $\rho < 1$. The last condition arises from the ESO parameters of serial sampling which are $v_i = \sigma_i\tau\|\mathbf{A}_i\|^2$; see Example 2. Finding optimal parameters is equivalent to equating the above inequalities. Note that the first two conditions (with equality) are equivalent to $\theta\bar{\tau} = (1-\theta)/2$ and $\bar{\sigma}_i = \frac{1-\theta}{2(p_i-(1-\theta))}$. With these choices, the third condition in (31) reads

$$(1-\theta)^2\kappa \leq 4\rho^2 p_i(p_i - (1-\theta)), \qquad i = 1, \dots, n\,. \tag{32}$$

It follows from (32) that with $\tilde{\kappa} = 1 + \kappa/\rho^2$ it holds that

$$\theta \geq \max_i 1 - \frac{2p_i}{1+\sqrt{\tilde{\kappa}_i}}. \tag{33}$$

Example 3 (serial uniform sampling). We first consider uniform sampling; i.e., every block is sampled with the same probability $p_i = 1/n$. Then it is easy to see that the smallest achievable rate is given by

$$\theta_{\text{uni}} = 1 - \frac{2}{n + n \max_j \sqrt{\tilde{\kappa}_j}} \tag{34}$$

and the step sizes become

$$\sigma_i = \frac{\mu_i^{-1}}{\max_j \sqrt{\tilde{\kappa}_j} - 1}, \quad \tau = \frac{\mu_g^{-1}}{n - 2 + n \max_j \sqrt{\tilde{\kappa}_j}}.$$

Example 4 (serial importance sampling). Instead of uniform sampling we may sample "important blocks" more often; i.e., we sample every block with a probability proportional to the square root of its condition number $p_i = \sqrt{\kappa_i} / \sum_j \sqrt{\kappa_j}$. Then the smallest rate that achieves (33) is given by

$$\theta_{\text{imp}} = 1 - \frac{2\nu}{\sum_{j=1}^n \sqrt{\kappa_j}}, \tag{35}$$

with $\nu \coloneqq \min_j \sqrt{\kappa_j}/(1 + \sqrt{\tilde{\kappa}_j})$, and the step sizes are

$$\sigma_i = \frac{\nu \mu_i^{-1}}{\sqrt{\kappa_i} - 2\nu}, \quad \tau = \frac{\nu \mu_g^{-1}}{\sum_{j=1}^n \sqrt{\kappa_j} - 2\nu}.$$

Example 5 (serial optimal sampling). Instead of a predefined probability we will look for an "optimal sampling" that minimizes the linear convergence rate θ. The optimal sampling can be found by equating condition (33) for $i = 1, \dots, n$,

$$\theta \left(1 + \sqrt{\tilde{\kappa}_i}\right) = 1 + \sqrt{\tilde{\kappa}_i} - 2p_i. \tag{36}$$

Summing (36) from 1 to n and using that for serial sampling $\sum_{i=1}^n p_i = 1$ leads to

$$\theta_{\text{opt}} = 1 - \frac{2}{n + \sum_{j=1}^n \sqrt{\tilde{\kappa}_j}}, \tag{37}$$

with step size parameters

$$\sigma_i = \frac{\mu_i^{-1}}{\sqrt{\tilde{\kappa}_i} - 1}, \quad \tau = \frac{\mu_g^{-1}}{n - 2 + \sum_{j=1}^n \sqrt{\tilde{\kappa}_j}}$$

and probabilities

$$p_i = \frac{1 + \sqrt{\tilde{\kappa}_i}}{n + \sum_{j=1}^n \sqrt{\tilde{\kappa}_j}}.$$

Remark 8 (minibatches). All arguments above can be readily extended to samplings where at each iteration not only one but a fixed number of blocks are chosen.

Remark 9 (better sampling). It is easy to see that optimal sampling is better than uniform sampling: if all condition numbers are the same, then the rates for uniform sampling (34) and optimal sampling (37) are equal, but if they are not, then the rate of optimal sampling is strictly smaller and thus better.

Moreover, optimal sampling is better than importance sampling. To see this, note that, due to the monotonicity of $\sqrt{x}/(1+\sqrt{1+x})$, we get

$$\begin{aligned}\theta_{\text{imp}} &= 1 - \min_i \frac{2}{\left(1+\sqrt{\tilde{\kappa}_i}\right)\sum_{j=1}^n \sqrt{\kappa_j/\rho^2}/\sqrt{\kappa_i/\rho^2}} \\ &\geq 1 - \min_i \frac{2}{\left(1+\sqrt{\tilde{\kappa}_i}\right)\sum_{j=1}^n (1+\sqrt{\tilde{\kappa}_j})/(1+\sqrt{\tilde{\kappa}_i})} = \theta_{\text{opt}}\,.\end{aligned}$$

Remark 10 (comparison to Zhang and Xiao [53]). The algorithm of Zhang and Xiao [53] is (almost[2]) a special case of the proposed algorithm where each block is picked with probability $p_i = 1/n$. Here m denotes the size of each block to be processed at every iteration, and n denotes the number of blocks. Moreover, they only consider the strongly convex case where g is μ_g-strongly convex and all f_i^* are μ_f-strongly convex. Then with R being the largest norm of the rows in $\mathbf{A}$, they achieve

$$\theta_{\text{ZX}} = 1 - \frac{1}{n + n\frac{\sqrt{m}R}{\sqrt{\mu_g\mu_f}}}\,.$$

If the minibatch size is $m = 1$, the blocks are chosen to be single rows, and the probabilities are uniform, then their rate is slightly worse than ours:

$$\begin{aligned}\theta_{\text{ZX}} = 1 - \frac{1}{n + n\max_j \sqrt{\kappa_j}} &\geq 1 - \frac{2}{2n + n\max_j \sqrt{\kappa_j/\rho^2}} \\ &\geq 1 - \frac{2}{2n + n(\max_j \sqrt{1+\kappa_j/\rho^2} - 1)} = \theta_{\text{uni}}\end{aligned}$$

for any $\rho \geq \frac{1}{2}$. For $m > 1$, the rates differ even more as the condition numbers are conservatively estimated. Similarly, the rates can be improved by nonuniform sampling if the row norms are not equal.

7. Numerical results. All numerical examples are implemented in Python using NumPy and the Operator Discretization Library (ODL) [1]. The python code and all example data are available from https://github.com/mehrhardt/spdhg.

7.1. Nonstrongly convex PET reconstruction. In this example we consider positron emission tomography (PET) reconstruction with a total variation (TV) prior. The goal in PET imaging is to reconstruct the distribution of a radioactive tracer from its line integrals [32]. Let $\mathbb{X} = \mathbb{R}^{d_1 \times d_2}$, $d_1 = d_2 = 250$, be the space of tracer distributions (images) and $\mathbb{Y}_i = \mathbb{R}^{|\mathbb{B}_i|}$ the data spaces where $\mathbb{B}_i \subset \{1,\dots,N\}$, $N = 200 \cdot 250$ (200 views around the object), are subsets of indices with $\mathbb{B}_i \cap \mathbb{B}_j = \emptyset$ if $i \neq j$ and $\cup_{i=1}^n \mathbb{B}_i = \{1,\dots,N\}$. All samplings in this example divide the views equidistantly. It is standard that PET reconstruction can be posed as the optimization problem (1), where the data fidelity term is given by the Kullback–Leibler divergence

$$f_i(y) = \begin{cases} \sum_{j\in\mathbb{B}_i} y_j + r_j - b_j + b_j \log\left(\frac{b_j}{y_j + r_j}\right) & \text{if } y_j + r_j > 0, \\ \infty & \text{else,} \end{cases} \tag{38}$$

[2] In contrast to our work, they have an extrapolation on both primal and dual variables. However, both extrapolations are related, as our extrapolation factor is the product of their extrapolation factors.

where it is convention that $0 \log 0 \coloneqq 0$. The operator $\mathbf{A}$ is a scaled X-ray transform where in each of 200 directions 250 line integrals are computed with the ASTRA toolbox [51, 50]. The prior is the TV of x with nonnegativity constraint, i.e., $g(x) = \alpha\|\nabla x\|_{2,1} + \iota_{\geq 0}(x)$, with regularization parameter $\alpha = 0.2$, and the gradient operator $\nabla x = (\nabla_1 x, \nabla_2 x) \in \mathbb{R}^{d_1 \cdot d_2 \times 2}$ is discretized by forward differences in horizontal and vertical direction, cf. [12] for details. The 2, 1-norm of these gradients is defined as $\|x\|_{2,1} \coloneqq \sum_j \sqrt{(\nabla_1 x_j)^2 + (\nabla_2 x_j)^2}$. The Fenchel conjugate of the Kullback–Leibler divergence (38) is

$$f_i^*(z) = \sum_{j \in \mathbb{B}_i} \begin{cases} -z_j r_j - b_j \log(1 - z_j) & \text{if } z_j \leq 1 \text{ and } (b_j = 0 \text{ or } z_j < 1), \\ \infty & \text{else,} \end{cases} \tag{39}$$

its proximal operator given by

$$\left[\operatorname{prox}_{f_i^*}^{\sigma_i}(z)\right]_j = \frac{1}{2}\left(z_j + 1 + \sigma_i r_j - \sqrt{(z_j - 1 + \sigma_i r_j)^2 + 4\sigma_i b_j}\right).$$

The proximal operator for g is approximated with 20 iterations of the fast gradient projection method (FGP) [6] with a warm start applied to the dual problem.

Parameters. In this experiment we choose $\gamma = 0.99$, $\theta = 1$, and all samplings are uniform, i.e., $p_i = 1/n$. The number of subsets varies among $n = 1$ (deterministic case), 50, and 250. The other step size parameters are chosen as

- PDHG, Pesquet and Repetti [35]: $\sigma_i = \tau = \gamma/\|\mathbf{A}\| \approx 6.9 \cdot 10^{-4}$;
- SPDHG: $\sigma_i = \gamma/\|\mathbf{A}_i\| \approx 2.2 \cdot 10^{-3}$, $\tau = \gamma/(n \max_i \|\mathbf{A}_i\|) \approx 2.2 \cdot 10^{-4}$.

Results. Figure 1 on the left shows that the ergodic Bregman distance converges with rate $1/k$, as proven in Theorem 4.3. On the right we compare the deterministic PDHG with the randomized SPDHG and the algorithm of Pesquet and Repetti. It can be clearly seen that the proposed SPDHG converges much faster than both the algorithm of Pesquet and Repetti and the deterministic PDHG. Some example images are found in Figure 2 after 5 epochs, which again highlights the speed-up gained by randomization.

7.2. TV denoising with Gaussian noise (primal acceleration). In the second example we consider denoising of an image that is degraded by Gaussian noise with the help of the anisotropic TV. This can be achieved by solving (1) with $\mathbb{X} = \mathbb{R}^{d_1 \times d_2}$, $d_1 = 442$, $d_2 = 331$; the data fit $g(x) = 1/(2\alpha)\|x - b\|_2^2$ is the squared Euclidean norm, and the prior is the (anisotropic) TV $f_i(y_i) = \|y_i\|_1, \mathbf{A}_i = \nabla_i$, and $n = 2$. Instead of the isotropic TV as in the previous example we consider here the anisotropic version, as it is separable in the direction of the gradient. The regularization parameter is chosen to be $\alpha = 0.12$. See, e.g., [13] for details on convex conjugates and proximal operators of these functionals.

Parameters. In this experiment we choose $\gamma = 0.99$ and the sampling to be uniform, i.e., $p_i = 1/n$. The number of subsets is either $n = 1$ in the deterministic case or $n = 2$ in the stochastic case. The (initial) step size parameters are

- PDHG, PA-PDHG, Pesquet and Repetti: $\sigma_i^{(0)} = \tau^{(0)} = \gamma/\|\mathbf{A}\| \approx 0.35$;
- SPDHG, PA-SPDHG: $\sigma_i^{(0)} = \gamma/\|\mathbf{A}_i\| \approx 0.50$, $\tau^{(0)} = \gamma/(n \max_i \|\mathbf{A}_i\|) \approx 0.25$.

The step sizes for acceleration vary with the iteration, with the primal step size $\tau_{(k)}$ getting smaller and the dual step size $\sigma^{(k)}$ getting larger. The extrapolation factor θ is chosen to be 1 for nonaccelerated, and converging to 1 for accelerated algorithms.

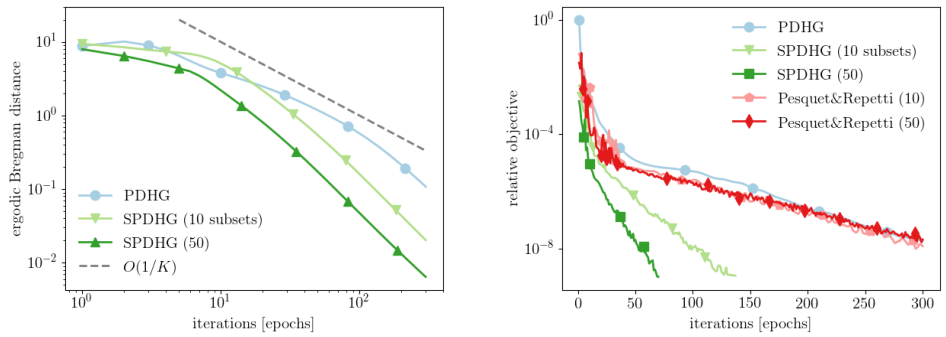

FIG. 1. *PET reconstruction with TV solved as a nonstrongly convex problem.* Left: *As proven in Theorem* 4.3*, the ergodic Bregman distances converge indeed with rate* $O(1/K)$. Right: *Speed comparison measured in terms of relative objective* $[\Phi(x^{(K)}) - \Phi(x^\sharp)]/[\Phi(x^{(0)}) - \Phi(x^\sharp)]$. *The proposed algorithm SPDHG converges faster than the algorithm of Pesquet and Repetti* [35] *and the deterministic PDHG.*

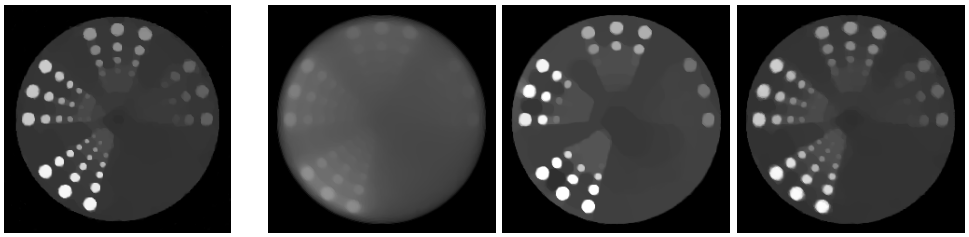

FIG. 2. *PET reconstruction results after* 5 *epochs with uniform sampling of* 50 *subsets.* From left to right: *Approximate primal part of saddle point, PDHG, Pesquet and Repetti* [35]*, and SPDHG. With the same number of operator evaluations, both stochastic algorithms make much more progress towards the saddle point.*

Results. The quantitative results in Figure 3 show that the accelerated algorithms are much faster than the nonaccelerated versions. Moreover, it can be seen that the stochastic variant of the accelerated PA-PDHG is even faster than its deterministic variant. In addition, the results show that the accelerated SPDHG indeed converges as $1/K^2$ in the norm of the primal part. Visual assessment of the denoised images in Figure 4 confirms these conclusions.

7.3. Huber-TV deblurring (dual acceleration). In the third example we consider deblurring with a known convolution kernel where the forward operator $\mathbf{A}_1$ resembles the convolution of images in $\mathbb{X} = \mathbb{R}^{d_1 \times d_2}, d_1 = 408, d_2 = 544$ with a motion blur of size 15×15. The noise is modeled to be Poisson with a constant background of 200 compared to the approximate data mean of 694.3. We further assume to have the knowledge that the reconstructed image should be nonnegative and upper-bounded by 100. By the nature of the forward operator, $\mathbf{A}x \geq 0$ whenever $x \geq 0$. Therefore the solution to (1) with the Kullback–Leibler divergence (38) remains the same if we replace the Kullback–Leibler divergence by the differentiable

$$f_1(y) = \sum_{i=1}^{N} \begin{cases} y_i + r_i - b_i + b_i \log\left(\frac{b_i}{y_i + r_i}\right) & \text{if } y_i \geq 0, \\ \frac{b_i}{2r_i^2} y_i^2 + \left(1 - \frac{b_i}{r_i}\right) y_i + r_i - b_i + b_i \log\left(\frac{b_i}{r_i}\right) & \text{else,} \end{cases} \tag{40}$$

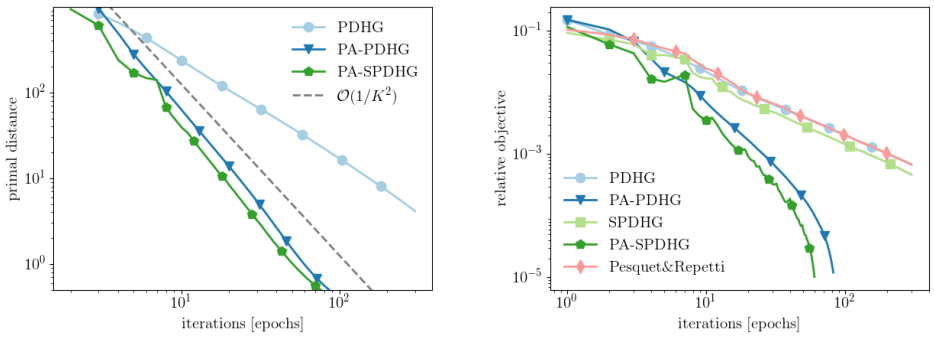

FIG. 3. *Primal acceleration for TV denoising.* Left: *Primal distance to saddle point* $\|x^{(K)} - x^\sharp\|^2$ Right: *Relative objective* $[\Phi(x^{(K)}) - \Phi(x^\sharp)]/[\Phi(x^{(0)}) - \Phi(x^\sharp)]$.

FIG. 4. *TV denoised images.* Left: *Approximate primal part of saddle point* $x^\sharp$ *after* 2,000 *PDHG iterations.* Right: *PDHG and primal accelerated SPDHG (PA-SPDHG) after* 20 *epochs.*

which has a $\left(\max_i b_i/r_i^2\right)$ Lipschitz continuous gradient. The Lipschitz constant is well-defined and nonzero as both the data b_i as well as the background r_i are positive. In our numerical example it is approximately 0.31.

Prior smoothness information is represented by the anisotropic TV with Huberized norm

$$f_i(\mathbf{A}_i x) = \alpha \sum_j \begin{cases} |y_j| & \text{if } |y_j| > \eta, \\ \frac{1}{2\eta}|y_j|^2 + \frac{\eta}{2} & \text{else} \end{cases}$$

for $i = 2, 3$, where $y_j = \nabla_{i-1} x_j$ are finite differences, $\eta = 1$, and the regularization parameter $\alpha = 0.1$. The constraints on the image are enforced by the indicator function $g = \imath_{\mathbb{B}}$ with $\mathbb{B} = \{x \in \mathbb{X} \,|\, 0 \le x_j \le 100\}$.

The convex conjugate of the modified Kullback–Leibler divergence (40) is

$$f_1^*(z) = \sum_{i=1}^N \begin{cases} \frac{r_i^2}{2b_i} z_i^2 + \left(r_i - \frac{r_i^2}{b_i}\right) z_i + \frac{r_i^2}{2b_i} + \frac{3b_i}{2} - 2r_i - b_i \log\left(\frac{b_i}{r_i}\right) & \text{if } z_i < 1 - \frac{b_i}{r_i}, \\ -r_i z_i - b_i \log(1 - z_i) & \text{if } 1 - \frac{b_i}{r_i} \le z_i < 1, \\ \infty & \text{if } z_i \ge 1, \end{cases}$$

which is $\left(\min_i r_i^2/b_i\right)$-strongly convex with proximal operator

$$\left[\operatorname{prox}_{f_1^*}^\sigma(z)\right]_i = \begin{cases} \frac{b_i z_i - \sigma r_i b_i + \sigma r_i^2}{b_i + \sigma r_i^2} & \text{if } z_i < 1 - \frac{b_i}{r_i}, \\ \frac{1}{2}\left\{z_i + \sigma r_i + 1 - \sqrt{(z_i + \sigma r_i - 1)^2 + 4\sigma b_i}\right\} & \text{else.} \end{cases}$$

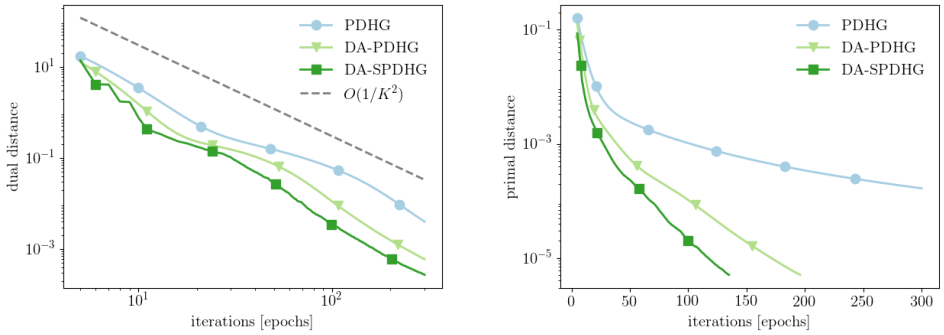

FIG. 5. *Dual acceleration for Huber-TV deblurring. Acceleration speeds up the convergence of both the dual variable (*left*) and the primal variable (*right*). Randomization in conjunction with acceleration yields even faster convergence. The accelerated algorithms converge with $O(1/K^2)$ in the dual distance (dashed line).*

FIG. 6. *Results after* 50 *epochs for deblurring with Huber-TV.* From left to right: *Blurry and noisy data with kernel (magnified), PDHG, and DA-SPDHG.*

Parameters. In this experiment we choose $\gamma = 0.99$ and consider uniform sampling, i.e., $p_i = 1/n$. The number of subsets is either $n = 1$ in the deterministic case or $n = 3$ in the stochastic case. The (initial) step size parameters are chosen to be

- PDHG: $\sigma_i = \tau = \gamma/\|\mathbf{A}\| \approx 0.095$;
- DA-PDHG: $\tilde{\sigma}_i^{(0)} = \mu_f/\|\mathbf{A}\| \approx 0.096$, $\tau^{(0)} = \gamma/\|\mathbf{A}\| \approx 0.095$;
- DA-SPDHG: $\tilde{\sigma}^{(0)} = \min_i \frac{\mu_i p_i^2}{\tau^0 \|\mathbf{A}_i\|^2 + 2\mu_i p_i (1-p_i)} \approx 0.073$,
  $\tau^{(0)} = 1/(n \max_i \|\mathbf{A}_i\|) \approx 0.032$.

Results. The quantitative results in Figure 5 show that the algorithm converges indeed with rate $O(1/K^2)$, as proven in Theorem 5.1. Moreover, they also show that randomization and acceleration can be used in conjunction for further speed-ups. The example images in Figure 6 show that randomization may lead to sharper images with the same number of epochs.

7.4. PET reconstruction (linear rate). For the final example we turn back to PET reconstruction, but this time with linear convergence rate. This means we want to solve the same minimization problem as in the first example, but now we replace the Kullback–Leibler functional with its modified version as in the previous

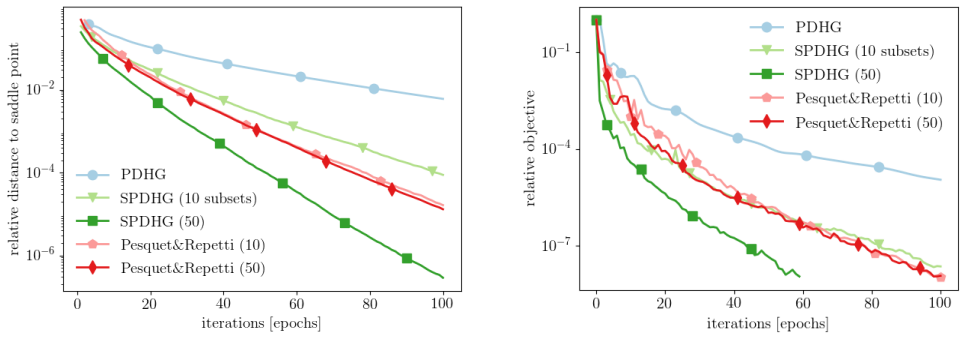

FIG. 7. *PET reconstruction with a strongly convex TV prior. Both the distance to the saddle point (*left*) and the objective value (*right*) show the speed-up by randomization over the deterministic PDHG. Moreover, for* 50 *subsets SPDHG is much faster than the algorithm proposed by Pesquet and Repetti. Also note the linear convergence on the left as proven in Theorem* 6.1.

example. We note again that this does not change the solution of the minimization problem. Moreover, to make TV strongly convex we add another regularization term, $\mu/2\|x\|_2^2$, to g. Note that the proximal operator of TV (indeed any functional) with added squared ℓ^2-norm, i.e., $g(x) = \alpha\,\mathrm{TV}(x) + \mu/2\|x\|_2^2$, can be solved by means of the original proximal operator $\mathrm{prox}_g^\sigma(z) = \mathrm{prox}_{\mathrm{TV}}^{\sigma\alpha/(1+\sigma\mu)}(z/(1+\sigma\mu))$. The regularization parameters are chosen as $\alpha = 0.05$ and $\mu = 0.5$.

Parameters. In this experiment we choose $\rho = 0.99$ and the sampling to be uniform, as the operators $\mathbf{A}_i$ all have similar norms. The step size parameters are chosen as derived in subsection 6.1; in particular, we choose

- PDHG: $\sigma \approx 3.8 \cdot 10^{-4}, \tau \approx 4.8 \cdot 10^{-3}, \theta \approx 0.995$;
- Pesquet and Repetti: $\sigma_i = \tau = \gamma/\|\mathbf{A}\| \approx 1.4 \cdot 10^{-3}$;
- SPDHG ($n = 10$ subsets): $\sigma_i \approx 1.2 \cdot 10^{-3}$, $\tau \approx 1.5 \cdot 10^{-3}$, $\theta^n \approx 0.985$;
- SPDHG ($n = 50$): $\sigma_i \approx 2.4 \cdot 10^{-3}$, $\tau \approx 5.8 \cdot 10^{-4}$, $\theta^n \approx 0.971$.

Note that the contraction rates of one epoch θ^n already indicate that SPDHG ($n = 50$) may be faster than PDHG and SPDHG ($n = 10$).

Results. The quantitative results in Figure 7 in terms of both distance to saddle point and objective value show that randomization speeds up the convergence so that both SPDHG and the algorithm of Pesquet and Repetti are faster than the deterministic PDHG. Interestingly, while more subsets make SPDHG faster, this does not hold for the algorithm of Pesquet and Repetti, where the speed seems to be constant with respect to the number of subsets. Moreover, the plot on the left confirms the linear convergence as proven in Theorem 6.1. The visual results in Figure 8 confirm these observations, as SPDHG with 50 subsets and 10 epochs is (in contrast to PDHG) visually already very close to the saddle point.

8. Conclusions and future work. We proposed a natural stochastic generalization of the deterministic PDHG algorithm to convex-concave saddle point problems that are separable in the dual variable. The analysis was carried out in the context of *arbitrary samplings*, which enabled us to obtain known deterministic convergence results as special cases. We proposed optimal choices of the step size parameters with which the proposed algorithm showed superior empirical performance on a variety of optimization problems in imaging.

In the future, we would like to extend the analysis to include iteration-dependent

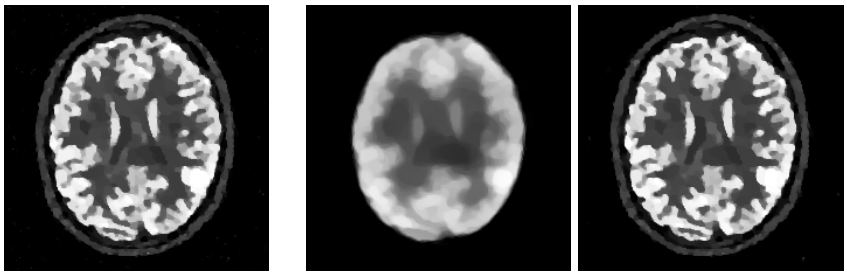

FIG. 8. *PET reconstruction results after* 10 *epochs.* Left: *Approximate primal part of saddle point computed by* 2,000 *iterations of PDHG.* Right: *PDHG and SPDHG (*50 *subsets).*

(adaptive) probabilities [17] and strong convexity parameters to further exploit the structure of many relevant problems. Moreover, the present optimal sampling strategies are only for scalar-valued step sizes and serial sampling. In the future, we wish to extend this to other sampling strategies such as multiblock or parallel sampling.

Appendix A. Postponed proofs.

Proof of Lemma 4.2. With the definition of y^+ and $\mathbf{Q}$ we have by completing the norm for any $x \in \mathbb{X}$ that

$$\begin{aligned} 2\langle \mathbf{Q}\mathbf{A}x, y^+ - y\rangle &= 2\left\langle x, \sum_{i\in\mathbb{S}} \mathbf{A}_i^* p_i^{-1}(\hat{y}_i - y_i)\right\rangle \\ &= 2\left\langle c^{1/2}\mathbf{T}^{-1/2}x, c^{-1/2}\sum_{i\in\mathbb{S}} \mathbf{C}_i^* p_i^{-1}\mathbf{S}_i^{-1/2}(\hat{y}_i - y_i)\right\rangle \\ &\geq -c\|x\|^2_{\mathbf{T}^{-1}} - \frac{1}{c}\left\|\sum_{i\in\mathbb{S}} \mathbf{C}_i^* z_i\right\|^2, \end{aligned} \tag{41}$$

where we used $z_i \coloneqq p_i^{-1}\mathbf{S}_i^{-1/2}(\hat{y}_i - y_i)$. Moreover, the expectation of the second term on the right-hand side of (41) can be estimated as

$$\mathbb{E}_{\mathbb{S}}\left\|\sum_{i\in\mathbb{S}} \mathbf{C}_i^* z_i\right\|^2 \leq \sum_{i=1}^n p_i v_i \|z_i\|^2 \leq \left(\max_i \frac{v_i}{p_i}\right)\sum_{i=1}^n p_i^2\|z_i\|^2, \tag{42}$$

where the first inequality is due to the ESO inequality (4). Inserting z leads to

$$\sum_{i=1}^n p_i^2\|z_i\|^2 = \sum_{i=1}^n p_i\|\hat{y}_i - y_i\|^2_{p_i^{-1}\mathbf{S}_i^{-1}} = \mathbb{E}_{\mathbb{S}}\|y^+ - y\|^2_{\mathbf{Q}\mathbf{S}^{-1}}, \tag{43}$$

where the last equation holds true by the definition of the expectation. Combining the expected value of inequality (41) with (42) and (43) yields the assertion. ∎

Proof of Lemma 4.4. By the definition of the proximal operator, for any $(x, y) \in \mathbb{W}$ it holds that

$$g(x) \geq g(x^{(k+1)}) + \langle \mathbf{T}_{(k)}^{-1}(x^{(k)} - x^{(k+1)}) - \mathbf{A}^*\overline{y}^{(k)}, x - x^{(k+1)}\rangle + \frac{\mu_g}{2}\|x - x^{(k+1)}\|^2,$$

$$f_i^*(y_i) \geq f_i^*(\hat{y}_i^{(k+1)}) + \langle (\mathbf{S}_{(k)}^i)^{-1}(y_i^{(k)} - \hat{y}_i^{(k+1)}) + \mathbf{A}_i x^{(k+1)}, y_i - \hat{y}_i^{(k+1)}\rangle + \frac{\mu_i}{2}\|y - \hat{y}^{(k+1)}\|^2$$

for $i = 1, \dots, n$. Summing twice all inequalities and exploiting the identity

$$2\langle \mathbf{B}(a-b), c-b\rangle = \|a-b\|_{\mathbf{B}}^2 + \|b-c\|_{\mathbf{B}}^2 - \|a-c\|_{\mathbf{B}}^2$$

yields

$$\begin{aligned}\|x^{(k)} - x\|^2_{\mathbf{T}_{(k)}^{-1}} + \|y^{(k)} - y\|^2_{\mathbf{S}_{(k)}^{-1}} \geq{}& \|x^{(k+1)} - x\|^2_{\mathbf{T}_{(k)}^{-1}+\mu_g\mathbf{I}} + \|\hat{y}^{(k+1)} - y\|^2_{\mathbf{S}_{(k)}^{-1}+\mathbf{M}} \\ &+ 2\left(g(x^{(k+1)}) - g(x) + f^*(\hat{y}^{(k+1)}) - f^*(y)\right) \\ &+ 2\left(\langle \mathbf{A}x^{(k+1)}, y - \hat{y}^{(k+1)}\rangle - \langle \mathbf{A}(x - x^{(k+1)}), \overline{y}^{(k)}\rangle\right) \\ &+ \|x^{(k+1)} - x^{(k)}\|^2_{\mathbf{T}_{(k)}^{-1}} + \|\hat{y}^{(k+1)} - y^{(k)}\|^2_{\mathbf{S}_{(k)}^{-1}},\end{aligned}$$

where we used the definition of the inner product and the norm on the product space $\mathbb{Y}$. It now suffices to complete the generalized distances $\mathcal{G}(x^{(k+1)}|w)$ and $\mathcal{F}(\hat{y}^{(k+1)}|w)$. ∎

Proof of Lemma 5.2. We follow a similar line of arguments as in the proof of Theorem 4.3. Note that for any saddle point $w^\sharp = (x^\sharp, y^\sharp)$ we have

$$2\mathcal{H}(x^{(k+1)}, \hat{y}^{(k+1)}|w^\sharp) = 2D_h^q(x^{(k+1)}, \hat{y}^{(k+1)}, w^\sharp) \geq \|x^{(k+1)} - x^\sharp\|^2_{\mu_g} + \|\hat{y}^{(k+1)} - y^\sharp\|^2_{\mathbf{M}}$$

such that the estimate of Lemma 4.4 can be written with $w = w^\sharp$ as

$$\begin{aligned}\|x^{(k)} - x^\sharp\|^2_{\mathbf{T}_{(k)}^{-1}} + \|y^{(k)} - y^\sharp\|^2_{\mathbf{S}_{(k)}^{-1}} \geq{}& \|x^{(k+1)} - x^\sharp\|^2_{\mathbf{T}_{(k)}^{-1}+2\mu_g\mathbf{I}} + \|\hat{y}^{(k+1)} - y^\sharp\|^2_{\mathbf{S}_{(k)}^{-1}+2\mathbf{M}} \\ &- 2\langle \mathbf{A}(x^{(k+1)} - x^\sharp), \hat{y}^{(k+1)} - \overline{y}^{(k)}\rangle \\ &+ \|x^{(k+1)} - x^{(k)}\|^2_{\mathbf{T}_{(k)}^{-1}} + \|\hat{y}^{(k+1)} - y^{(k)}\|^2_{\mathbf{S}_{(k)}^{-1}}\end{aligned}$$

using the rule (14). With (11) and (12) and again (14) we arrive at

$$\begin{aligned}&\|x^{(k)} - x^\sharp\|^2_{\mathbf{T}_{(k)}^{-1}} + \|y^{(k)} - y^\sharp\|^2_{\mathbf{Q}\mathbf{S}_{(k)}^{-1}+2\mathbf{M}(\mathbf{Q}-\mathbf{I})} \\ &\geq \mathbb{E}^{(k+1)}\Big\{\|x^{(k+1)} - x^\sharp\|^2_{\mathbf{T}_{(k)}^{-1}+2\mu_g\mathbf{I}} + \|\hat{y}^{(k+1)} - y^\sharp\|^2_{\mathbf{Q}\mathbf{S}_{(k)}^{-1}+2\mathbf{M}\mathbf{Q}} \\ &\quad - 2\langle \mathbf{A}(x^{(k+1)} - x^\sharp), \mathbf{Q}(y^{(k+1)} - y^{(k)}) + y^{(k)} - \overline{y}^{(k)}\rangle \\ &\quad + \|x^{(k+1)} - x^{(k)}\|^2_{\mathbf{T}_{(k)}^{-1}} + \|y^{(k+1)} - y^{(k)}\|^2_{\mathbf{Q}\mathbf{S}_{(k)}^{-1}}\Big\}. \qquad (44)\end{aligned}$$

Inserting the extrapolation (22) into the inner product yields

$$\begin{aligned}&-\theta_{(k-1)}\langle \mathbf{Q}\mathbf{A}(x^{(k)} - x^\sharp), y^{(k)} - y^{(k-1)}\rangle \\ &\geq \mathbb{E}^{(k+1)}\Big\{\theta_{(k-1)}\langle \mathbf{Q}\mathbf{A}(x^{(k+1)} - x^{(k)}), y^{(k)} - y^{(k-1)}\rangle \\ &\quad - \langle \mathbf{Q}\mathbf{A}(x^{(k+1)} - x^\sharp), y^{(k+1)} - y^{(k)}\rangle\Big\}. \qquad (45)\end{aligned}$$

The assertion is shown by taking the expectations $\mathbb{E}^{(k,k-1)} := \mathbb{E}^{(k)}\mathbb{E}^{(k-1)}$ on (44), using (45), and estimating the last inner product by Lemma 4.2 as

$$\begin{aligned}&2\theta_{(k-1)}\mathbb{E}^{(k+1)}\langle \mathbf{Q}\mathbf{A}(x^{(k+1)} - x^{(k)}), y^{(k)} - y^{(k-1)}\rangle \\ &\geq -\mathbb{E}^{(k+1)}\left\{\|x^{(k+1)} - x^{(k)}\|^2_{\mathbf{T}_{(k)}^{-1}} + (\gamma\theta_{(k-1)})^2\|y^{(k)} - y^{(k-1)}\|^2_{\mathbf{Q}\mathbf{S}_{(k)}^{-1}}\right\}.\end{aligned}$$

∎

REFERENCES

[1] J. Adler, H. Kohr, and O. Öktem, *Operator Discretization Library (ODL)*, 2017, https://github.com/odlgroup/odl.

[2] Z. Allen-Zhu, Y. Yuan, P. Richtárik, and Y. Yuan, *Even faster accelerated coordinate descent using non-uniform sampling*, in International Conference on Machine Learning, Proc. Mach. Learn. Res. 48, 2016; preprint available from https://arxiv.org/abs/1512.09103.

[3] P. Balamurugan and F. Bach, *Stochastic variance reduction methods for saddle-point problems*, in Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16), 2016, pp. 1416–1424.

[4] H. Bauschke and J. Borwein, *Legendre functions and the method of random Bregman projections*, J. Convex Anal., 4 (1997), pp. 27–67.

[5] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, 2011, https://doi.org/10.1007/978-1-4419-9467-7.

[6] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202, https://doi.org/10.1137/080716542.

[7] M. Benning, C.-B. Schönlieb, T. Valkonen, and V. Vlačić, *Explorations on Anisotropic Regularisation of Dynamic Inverse Problems by Bilevel Optimisation*, preprint, https://arxiv.org/abs/1602.01278, 2016.

[8] D. P. Bertsekas, *Incremental Gradient, Subgradient, and Proximal Methods for Convex Optimization: A Survey*, in Optimization for Machine Learning, S. Sra, S. Nowozin, and S. J. Wright, eds., MIT Press, 2011, pp. 85–120.

[9] D. P. Bertsekas, *Incremental proximal methods for large scale convex optimization*, Math. Program., 129 (2011), pp. 163–195, https://doi.org/10.1007/s10107-011-0472-0.

[10] D. Blatt, A. O. Hero, and H. Gauchman, *A convergent incremental gradient method with a constant step size*, SIAM J. Optim., 18 (2007), pp. 29–51, https://doi.org/10.1137/040615961.

[11] K. Bredies and M. Holler, *A TGV-based framework for variational image decompression, zooming, and reconstruction. Part* I: *Analytics*, SIAM J. Imaging Sci., 8 (2015), pp. 2814–2850, https://doi.org/10.1137/15M1023865.

[12] A. Chambolle, *An algorithm for total variation minimization and applications*, J. Math. Imaging Vision, 20 (2004), pp. 89–97.

[13] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vision, 40 (2011), pp. 120–145, https://doi.org/10.1007/s10851-010-0251-1.

[14] A. Chambolle and T. Pock, *An introduction to continuous optimization for imaging*, Acta Numer., 25 (2016), pp. 161–319, https://doi.org/10.1017/S096249291600009X.

[15] A. Chambolle and T. Pock, *On the ergodic convergence rates of a first-order primal-dual algorithm*, Math. Program., 159 (2016), pp. 253–287, https://doi.org/10.1007/s10107-015-0957-3.

[16] P. L. Combettes and J.-C. Pesquet, *Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping*, SIAM J. Optim., 25 (2015), pp. 1221–1248, https://doi.org/10.1137/140971233.

[17] D. Csiba, Z. Qu, and P. Richtárik, *Stochastic dual coordinate ascent with adaptive probabilities*, J. Mach. Learn. Res., 37 (2015), pp. 674–683.

[18] C. D. Dang and G. Lan, *Randomized Methods for Saddle Point Computation*, preprint, https://arxiv.org/abs/1409.8625, 2014.

[19] A. Defazio, F. Bach, and S. Lacoste-Julien, *SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives*, in Advances in Neural Information Processing Systems 27 (NIPS 2014), Curran Associates, 2014, pp. 1646–1654, https://arxiv.org/abs/1407.0202v2.

[20] R. M. de Oliviera, E. S. Helou, and E. F. Costa, *String-averaging incremental subgradients for constrained convex optimization with applications to reconstruction of tomographic images*, Inverse Problems, 32 (2016), 115014, https://doi.org/10.1088/0266-5611/32/11/115014.

[21] E. Esser, X. Zhang, and T. F. Chan, *A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science*, SIAM J. Imaging Sci., 3 (2010), pp. 1015–1046, https://doi.org/10.1137/09076934X.

[22] V. Estellers, S. Soatto, and X. Bresson, *Adaptive regularization with the structure tensor*, IEEE Trans. Image Process., 24 (2015), pp. 1777–1790, https://doi.org/10.1109/TIP.2015.2409562.

[23] O. Fercoq and P. Bianchi, *A Coordinate Descent Primal-Dual Algorithm with Large Step Size and Possibly Non Separable Functions*, preprint, https://arxiv.org/abs/1508.04625, 2015.
[24] O. Fercoq and P. Richtárik, *Accelerated, parallel, and proximal coordinate descent*, SIAM J. Optim., 25 (2015), pp. 1997–2023, https://doi.org/10.1137/130949993.
[25] X. Gao, Y. Xu, and S. Zhang, *Randomized Primal-Dual Proximal Block Coordinate Updates*, preprint, https://arxiv.org/abs/1605.05969, 2016.
[26] G. Gilboa, M. Moeller, and M. Burger, *Nonlinear spectral analysis via one-homogeneous functionals: Overview and future prospects*, J. Math. Imaging Vis., 56 (2016), pp. 300–319, https://doi.org/10.1007/s10851-016-0665-5.
[27] F. Knoll, M. Holler, T. Koesters, R. Otazo, K. Bredies, and D. K. Sodickson, *Joint MR-PET reconstruction using a multi-channel image regularizer*, IEEE Trans. Medical Imaging, 36 (2017), pp. 1–16, https://doi.org/10.1109/TMI.2016.2564989.
[28] J. Konečný, J. Liu, P. Richtárik, and M. Takáč, *Mini-batch semi-stochastic gradient descent in the proximal setting*, IEEE J. Selected Topics Signal Process., 10 (2016), pp. 242–255.
[29] R. D. Kongskov, Y. Dong, and K. Knudsen, *Directional Total Generalized Variation Regularization*, preprint, https://arxiv.org/abs/1701.02675, 2017.
[30] P.-L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964–979, https://doi.org/10.1137/0716071.
[31] A. Nedić and D. P. Bertsekas, *Incremental subgradient methods for nondifferentiable optimization*, SIAM J. Optim., 12 (2001), pp. 109–138, https://doi.org/10.1137/S1052623499362111.
[32] J. M. Ollinger and J. A. Fessler, *Positron emission tomography*, IEEE Signal Process. Mag., 14 (1997), pp. 43–55, https://doi.org/10.1109/79.560323.
[33] N. Parikh and S. P. Boyd, *Proximal algorithms*, Found. Trends Optim., 1 (2014), pp. 123–231, https://doi.org/10.1561/2400000003.
[34] Z. Peng, T. Wu, Y. Xu, M. Yan, and W. Yin, *Coordinate friendly structures, algorithms and applications*, Ann. Math. Sci. Appl., 1 (2016), pp. 1–54, https://doi.org/10.4310/AMSA.2016.v1.n1.a2.
[35] J.-C. Pesquet and A. Repetti, *A Class of Randomized Primal-Dual Algorithms for Distributed Optimization*, preprint, https://arxiv.org/abs/1406.6404, 2015.
[36] T. Pock and A. Chambolle, *Diagonal preconditioning for first order primal-dual algorithms in convex optimization*, in Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 1762–1769, https://doi.org/10.1109/ICCV.2011.6126441.
[37] T. Pock, D. Cremers, H. Bischof, and A. Chambolle, *An algorithm for minimizing the Mumford-Shah functional*, in Proceedings of the IEEE International Conference on Computer Vision, 2009, pp. 1133–1140, https://doi.org/10.1109/ICCV.2009.5459348.
[38] Z. Qu and P. Richtárik, *Coordinate descent with arbitrary sampling* I: *Algorithms and complexity*, Optim. Methods Softw., 31 (2016), pp. 829–857, https://doi.org/10.1080/10556788.2016.1190360.
[39] Z. Qu and P. Richtárik, *Quartz: Randomized dual coordinate ascent with arbitrary sampling*, in Proceedings of the 28th International Conference on Neural Information Processing Systems, Volume 1, 2015, pp. 865–873.
[40] P. Richtárik and M. Takáč, *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function*, Math. Program., 144 (2014), pp. 1–38, https://doi.org/10.1007/s10107-012-0614-z.
[41] P. Richtárik and M. Takáč, *On optimal probabilities in stochastic coordinate descent methods*, Optim. Lett., 10 (2016), pp. 1233–1243, https://doi.org/10.1007/s11590-015-0916-1.
[42] P. Richtárik and M. Takáč, *Parallel coordinate descent methods for big data optimization*, Math. Program., 156 (2016), pp. 433–484.
[43] D. Rigie and P. La Riviere, *Joint reconstruction of multi-channel, spectral CT data via constrained total nuclear variation minimization*, Phys. Med. Biol., 60 (2015), pp. 1741–1762, https://doi.org/10.1088/0031-9155/60/4/1741.
[44] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.
[45] L. Rosasco and S. Villa, *Stochastic Inertial Primal-Dual Algorithms*, preprint, https://arxiv.org/abs/1507.00852v1, 2015.
[46] M. Schmidt, N. Le Roux, and F. Bach, *Minimizing finite sums with the stochastic average gradient*, Math. Program., 162 (2017), pp. 83–112, https://doi.org/10.1007/s10107-016-1030-6.
[47] M. Takác, A. Bijral, P. Richtárik, and N. Srebro, *Mini-batch primal and dual methods for SVMs*, in Proceedings of the 30th International Conference on Machine Learning, Springer, 2013, pp. 537–552.

[48] P. Tseng, *An incremental gradient(-projection) method with momentum term and adaptive stepsize rule*, SIAM J. Optim., 8 (1998), pp. 506–531, https://doi.org/10.1137/S1052623495294797.

[49] T. Valkonen, *Block-Proximal Methods with Spatially Adapted Acceleration*, preprint, https://arxiv.org/abs/1609.07373, 2016.

[50] W. van Aarle, W. J. Palenstijn, J. Cant, E. Janssens, F. Bleichrodt, A. Dabravolski, J. De Beenhouwer, K. Joost Batenburg, and J. Sijbers, *Fast and flexible X-ray tomography using the ASTRA toolbox*, Optics Express, 24 (2016), pp. 25129–25147, https://doi.org/10.1364/OE.24.025129.

[51] W. van Aarle, W. J. Palenstijn, J. De Beenhouwer, T. Altantzis, S. Bals, K. J. Batenburg, and J. Sijbers, *The ASTRA toolbox: A platform for advanced algorithm development in electron tomography*, Ultramicroscopy, 157 (2015), pp. 35–47, https://doi.org/10.1016/j.ultramic.2015.05.002.

[52] M. Wen, S. Yue, Y. Tan, and J. Peng, *A Randomized Inertial Primal-Dual Fixed Point Algorithm for Monotone Inclusions*, preprint, https://arxiv.org/abs/1611.05142, 2016.

[53] Y. Zhang and L. Xiao, *Stochastic primal-dual coordinate method for regularized empirical risk minimization*, in Proceedings of the 32nd International Conference on Machine Learning, 2015, pp. 1–34.

[54] L. W. Zhong and J. T. Kwok, *Fast stochastic alternating direction method of multipliers*, J. Mach. Learn. Res., 32 (2014), pp. 46–54.

[55] Z. Zhu and A. J. Storkey, *Adaptive stochastic primal-dual coordinate descent for separable saddle point problems*, in Machine Learning and Knowledge Discovery in Databases, A. Appice, P. P. Rodrigues, V. Santos Costa, C. Soares, J. Gama, and A. Jorge, eds., Springer, 2015, pp. 643–657.